Optimal Timing of Phase Resolved Cell Cycle Progression

With Applications to Immunology and Cell Culture Experiments

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

d o c t o r r e r u m n a t u r a l i u m
(Dr. Rer. Nat.)

im Fach Biologie

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät I
Humboldt-Universität zu Berlin

von
Dipl.-Phys. Tom S. Weber

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät I:
Prof. Stefan Hecht, PhD

Gutachter:
1. Dr. Michal Or-Guil
2. Dr. Jorge Carneiro
3. Prof. Dr. Carsten Hartmann

Tag der mündlichen Prüfung: 27.06.2014

à $\mathcal{E}lena$ ...

Contents

0 Introduction

Cell division, which produces two identical sister cells from a single parental cell, is one of the most fundamental processes in biology. While in single-celled organisms, cell division represents primarily a way to achieve self-reproduction, and is thus directly related to the organism's fitness, the role of cell division in the multicellular case is more multifaceted. It plays an essential role in as divergent processes as embryogenesis, innate and adaptive immunity, maintenance of tissues like the skin, cell differentiation, hematopoiesis, body growth, oogenesis and spermatogenesis. At the same time, deregulation of cell division, such as excessive growth but also insufficient proliferation, represent a major threat to the organism integrity and survival [1].

0.1 General background

The whole process of cell division is called the cell cycle. The cell cycle comprises a series of phases, allowing a cell to generate, from itself, two viable copies, which are able to repeat the same process again. In eukaryotic cells, there are four major cell cycle phases, which follow each other in a specific order: the G_1 phase, the S phase, the G_2 phase, and the M phase [2]. During the S phase, each pair of chromosomes of a cell is copied once [3], while during the M phase, the original and the newly replicated DNA molecules are partitioned into two, such that, at the time of division, each daughter cell receives an identical set of chromosomes [4]. At the end of the M phase, the cell engages into cytokinesis, which results in the physical separation of the two sister cells. The G_1 phase and the G_2 phase represent temporal gaps in between cytokinesis and the initiation of the S phase, and between the completion of the S phase and the initiation of the M phase, respectively. After the completion of cytokinesis, both sister cells are by definition in the G_1 phase. Whether new-born cells in the G_1 re-enter the S phase for another round of replication depends on external and internal signals [1]. Cells which do not re-initiate the S phase, can enter a quiescent phase instead, termed the G_0 phase [5] or undergo apoptosis.
Re-entry into the S phase is regulated by the so-called restriction point, a cell cycle checkpoint which ensures that a cell initiates DNA synthesis only if conditions are favorable [1]. Other cell cycle checkpoints exist which control whether the genome has been fully replicated [6] or whether chromosomes are properly aligned on the mitotic spindle [7]. The latter condition is a prerequisite for symmetric segregation of genomic information during and at the end of mitosis.
The cell cycle, and each of its phases, require a certain amount of time for their completion. This time is denoted the division time for the whole cell cycle, and phase completion time for the individual phases. Division and phase completion times are important characteristics of dividing cell populations, as they reflect, for instance, the time it takes cells to duplicate their genome, to properly segregate their chromosomes, to pass cell cycle checkpoints, to grow, or to make fate decisions concerning division and quiescence. Because the underlying processes are stochastic, division and phase completion times are stochastic as well. Hence, division and phase completion times are random variables with associated probability density functions rather than single values. In order to retrieve statistics of these densities from experimental data, appropriate mathematical techniques are required. Developing these techniques will become one of the main focuses in this thesis.
In the field of oncology, not the division or the phase completion times, but a related quantity, namely the potential doubling time T_{pot}, is often of interest. This quantity can directly be computed from the division and phase completion time densities. The opposite is however not true. If the division

and phase completion time densities are unknown, *ad hoc* assumptions about their shape are required in order to estimate T_{pot} from experimental data [8]. The impact of these assumptions on the final estimates can however be significant.

There exist several well established methods to measure division and phase completion times *in vivo* and *in vitro*. Dividing cells can be pulse labeled with nucleoside analogs [9–11]. The latter are incorporated into newly synthesizing DNA of cells which are in S phase during the pulse. By measuring, after the pulse, labeled cells as they enter and exit mitosis once or even twice, the phase completion times and the division times can be estimated. This technique is called the fraction of labeled mitoses (FLM) method [9, 11]. A more direct method to measure division times is by taking, using a light microscope, images of *in vitro* dividing cell populations at short time intervals. By simply counting the number of intervals that separate birth and division events of individual cells, division times are retrieved [12].

Nowadays, the FLM method has been replaced by a less tedious and more sophisticated technique, called DNA-BrdU pulse-chase labeling [13–16]. The latter method is based on a combination of nucleoside analogs (e.g., Bromodeoxyuridine (BrdU)) pulse labeling, DNA staining, and flow cytometry. This technique will be denoted in the following, for the sake of generality, the DNA-NA pulse-chase labeling method, where NA stands for one of several available nucleoside analogs (e.g., BrdU, IdU, EdU, CldU). As it is the case for the FLM method, the DNA-NA pulse-chase labeling technique first labels permanently cells in S phase with a given nucleoside analog, and then follows these cells as they progress through the subsequent phases. The DNA-NA pulse-chase labeling technique allows to measure the frequency of cells in G_0/G_1 phase (i.e., cells in either G_0 or G_1 phase), the S phase and the G_2/M phase (i.e., cells in either G_2 or M phase). In addition, the kinetics of four cell populations can be followed over time: 1) the fraction of labeled undivided cells, 2) the fraction of labeled divided cells, 3) the fraction of unlabeled undivided cells initially in G_2/M phase, and 4) the remaining cells [14, 17]. Moreover the relative movement of the labeled cell populations on the DNA axis can be estimated [15, 17].

Several approaches have been employed in the past to interpret DNA-NA pulse-chase labeling experiments [18–21]. While most rely on some sort of model-based analysis, basic information can be extracted from the respective data sets without referring to complex mathematical models. For example, given that the proliferating cell population is under homeostatic control and that the total cell count remains approximately constant over time, the average relative time cells spend in the G_0/G_1 phase, the S phase and the G_2/M phase is proportional to the frequency of cells in each of these phases. This however is not true for growing populations. A second basic quantity, which can be directly estimated for both growing and non-growing populations, is the minimal time it takes cells to complete the G_2/M phase. This equals the time elapsed between the pulse and the appearance of the first labeled divided cells.

Inferring, with the DNA-NA pulse-chase labeling method, more complex features from a dividing cell population, requires however more elaborate model assumptions. Typically cell cycle progression is specified in more detail by explicitly or implicitly defining the completion times for individual phases [18–21]. This allows to make theoretical predictions about the kinetics of labeled and unlabeled cell cohorts and about the relative movement of the labeled cell populations [14, 15]. Comparing these predictions with data from DNA-NA pulse-chase labeling experiments yields cell cycle progression estimates, for instance the average S phase completion time or the division time. In contrast to the basic quantities discussed before, absolute and not relative completion times are derived. Moreover, some information about the completion time distributions can be obtained [10, 19, 22]. This clearly motivates a model-based approach for the interpretation of DNA-NA pulse-chase labeling experiments, because important features of cell cycle progression can be inferred which are not easily accessible through other means.

Irrespectively however, any conclusions based on DNA-NA pulse-chase labeling experiments about a dividing cell population remain conditioned on the underlying assumptions about cell proliferation.

This is a consequence of the fact that the DNA-NA pulse-chase labeling technique is an indirect method. Individual cells are not observed as they progress through the various cell cycle phases, but snapshots of dividing cell populations are recorded at certain time points after pulse labeling, and the number of labeled and unlabeled cells in the different phases are counted. Such an approach harbors advantages and disadvantages compared to continuous observations of individual cells, the major advantage being its experimental feasibility, as direct observation of single cells remains, despite recent advances in the field of live imaging and video microscopy [23], extremely challenging, if not impossible. Especially, for *in vivo* experiments, quasi-continuous observations are currently restricted to relatively slow-migrating cells in superficial tissues with imaging periods of less than three hours [24–26], which is much shorter than the typical division time [27]. Moreover, in order to measure cell cycle phase durations and their variability by video microscopy, it would become necessary to identify cells in each of the cell cycle phases. At present however, the most widely accepted technique, namely the fluorescent ubiquitination-based cell cycle indicator (FUCCI), allows identification of cells in the aggregated phases G_0/G_1 and $S/G_2/M$ only [28, 29].

0.2 Objectives

The DNA-NA pulse-chase labeling method in combination with model-based data analysis represents a powerful technique to study cell proliferation *in vitro* and *in vivo*. It allows to estimate, besides of the phase completion times, clinically relevant quantities, like the potential doubling time [8, 30–32] and the S phase completion time of dividing cell populations. Currently however, there are several short-comings, which make this method only partly satisfactory. First, from a statistician's point of view, there has been no thorough study, concerned with the uncertainty in the estimates the method provides. This might be partly due to the circumstance, that no general analytical solutions have been derived so far for the proposed models. Secondly, the quantities that are estimated by this method represent often (although not exclusively [10, 19, 21]) mean values, and no information about the biological variance in cell cycle progression is provided. This information, while important in its own right, is in addition required to infer accurately T_{pot} or the division time. Thirdly, from an experimentalist's point of view, no clear guide-lines are available which dictate how to set up an DNA-NA pulse-chase labeling experiment most efficiently. While detailed protocols, describing working solutions for pulse labeling and staining procedures, are frequent in the literature [14–18, 33], the problem of choosing the time points at which samples should best be collected has not been systematically addressed so far. Finally, most studies which have interpreted DNA-NA pulse-chase labeling based their inference on a minimal set of assumptions about the cell cycle and the dividing cells [15, 17–21]. While these models might characterize well *in vitro* conditions, *in vivo* the complexity of cell cycle progression is likely to be much higher, since cell division is inter-wined with other processes. For instance, heterogeneity, cell migration and differentiation are ubiquitous factors, which can potentially confound inference results.

In this thesis, the aim is to develop a general mathematical framework, which should allow to estimate accurately phase completion times from DNA-NA pulse-chase labeling experiments. To achieve this goal, in a first step, we will develop a theoretical model of cell cycle progression, which will be sufficiently complex to describe DNA-NA pulse-chase labeling experiments, yet at the same time remain sufficiently abstract to be amenable for a mathematical treatment. Subsequently, we will derive analytical solutions and validate this model with *in vitro* DNA-NA pulse-chase labeling data from homogeneous exponentially growing cell cultures. Cell cycle progression parameters will be extracted using non-linear least-squares fitting. In order to estimate the impact of measurement error on the uncertainty in the inferred parameter values, we will then extend the model to the Bayesian inference framework. We will compute, for two data sets and two different cell lines, credibility intervals for each parameter of the model, as well as for the average phase durations and the division time. With the extended model, we will address the problem of how to choose optimal sampling schemes

which minimize the uncertainty in parameter estimates. Finally, we will adapt the model, in order to account for homeostatic control, heterogeneity and cell migration in dividing cell populations. This more complex model will be used to interpret *in vivo* cell cycle and migration data from germinal centers, which are important anatomical structures of the humoral adaptive immune system [34].

0.3 Outline

This thesis is divided into four chapters. Each of the first three chapters is further subdivided into three sections. While the first section of each of these chapters gives a brief motivation and some background information specific to the questions addressed therein, the second section contains the main results, and the third section provides a chapter specific discussion. The fourth and last chapter comprises a summary and a general discussion, which reviews, under a more general light, main assumptions and results. Finally, future directions and prospective applications are explored in the outlook section.

In **Chapter 1**, we focus on modeling basic cell cycle progression. A stochastic cell cycle model is developed, which assumes that the completion times for the G_1, the S and the G_2/M phase are given by three independent random variable. This model is analyzed on theoretical grounds, in order to understand not only its asymptotic behavior, but also to predict the outcomes of DNA-NA pulse-chase labeling experiments. Analytical average kinetics are then used to extract information about cell cycle parameters from DNA-NA pulse-chase labeling data under homogeneous *in vitro* conditions. Besides of non-linear least-squares fitting, Bayesian inference is applied to assess the uncertainty in the obtained parameter values. A special interest lies in estimating accurately the variabilities in the phase completion times, which, compared to the average times, are much more difficult to assess.

In **Chapter 2**, the improvement of experimental design of DNA-NA pulse-chase labeling experiments is addressed. Two different but complementary approaches are explored: The first approach employs the theory of D-optimal experimental design, to find the experimental schedules which are expected to reduce most the uncertainty in the parameter estimates. The second approach analyzes the potential of the double labeling technique. Instead of using a single pulse of nucleoside analogs, two pulses with two different nucleoside analogs are applied, and the performance of such a technique is tested *in silico*.

In **Chapter 3**, cell division in germinal centers is analyzed. The simple model, developed before, is adapted to interpret data from this more complex *in vivo* scenario. Heterogeneity and cell cycle dependent migration are included into the model, and five different independent data sets are used to test the model assumptions.

Finally, in **Chapter 4**, we recapitulate the contributions of this work, we discuss the assumptions underlying our modeling approach, we critically review its limitations, and consider briefly hypothetical applications of the proposed methodology to cancer growth prediction and treatment design.

1 A Stochastic Phase Resolved Cell Cycle Model

1.1 Motivation and Background

Due to the global dependency of the mammalian organism (including humans) on cell division, minute characterization and quantification of cell cycle progression can provide valuable insights into the regulation and deregulation of many processes related to health and disease. Direct and indirect experimentally determined properties are numerous and include the growth fraction (e.g., [35]), the potential doubling time (e.g., [20]), the cell division rate (e.g., [36]), the fraction of cells in the G_1, S, or G_2M phase (e.g., [37]), the mitotic index (e.g., [38]), the extent of CFSE dilution (e.g., [39]), cell cycle related gene expression (e.g., [40]), the growth curve (e.g., [41]), the division time (e.g., [42]), and long-term BrdU incorporation (e.g., [43]), among others.

In this chapter, we will develop a mathematical framework, to model and analyze stochastic phase-specific cell cycle progression. This will allow us to interpret data from a well known experimental technique, frequently called BrdU pulse-labeling or simply BrdU labeling [17], but termed here and in the following, for the sake of generality, the DNA-NA pulse-chase labeling method. As mentioned in the Introduction, the NA stands for any of the known nucleoside analogs, including BrdU. The latter method, if combined with an appropriate model-based inference strategy, turns out to be more informative than most other indirect approaches enumerated above. Crucial for this inference process is the underlying cell cycle model, which will be our main focus herein.

In the following, important characteristics of the cell cycle are briefly summarized, with a bias towards a macroscopic perspective, leaving for the vast field of molecular biology of the cell cycle relative little space. This reflects the level of description chosen in this thesis and corresponds consequently to the resolution of our modeling approach. The latter is characterized by reducing complex biological processes involving thousands of genes, billions of proteins (7.9×10^9, [2]) and trillions of molecules to a unique feature, namely their duration, or more precisely the distribution over their duration. This degree of abstraction allows to understand and predict certain kinetics observed in experimental data, with the caveat that the results remain largely unaffected by the specific underlying molecular or genetic interactions. Common techniques used to study cell cycle kinetics are then extensively listed. Finally, to conclude the background section, previous approaches to cell cycle modeling are reviewed. In the result section, a stochastic phase resolved cell cycle model is developed and analytical predictions for the average kinetics are derived. This model is then used to analyze data from two *in vitro* DNA-BrdU pulse-chase labeling experiments, with a strong theoretical and computational emphasis on parameter identification, revealing the full potential of a model-based approach. At the end of this chapter, limitations and possible improvements of the methodology are discussed.

1.1.1 Characteristics of the cell cycle

The most obvious function of the cell cycle is to generate, from a single parental cell, two identical daughter cells that are capable to repeat the same process again. From a naive technical point of view this 'task' poses a series of 'challenges'. In order to maintain the organism's genome intact, billions of DNA base pairs (e.g., 3 billion base pairs for the human genome [44]) have to be copied with high fidelity each time a cell divides [45], the chromosomes have to be segregated such that each daughter cell receives a complete genome [4], the differentiation state has to be passed from the parental to the daughter cells [46], the chromosomes have to be reproduced once and only once [3], and finally both

the cell size and the number of organelles have to be synchronized with cell division to ensure viability and functionality of the progeny [47].
Apparently, cells in a human sized organism master these tasks millions of times per day, without any major incidences. The 'secret' behind this smooth functioning, as has been shown through decades of research, is not found in error-free cell cycle progression, but lies in the tight control that acts at several points during the cell cycle [48]. Each time an irregularity is detected, cell cycle progression is halted, and appropriate repair mechanism are activated in order to resolve the 'problem'. If this is not possible, apoptotic pathways are initiated that eventually lead to the clearance of the defective cell [49].
A scaffold for this highly complex control system is provided by the various phases of the cell cycle, which will be briefly described in the following section.

The cell cycle phases

The cell cycle is a sequence of processes that can be grouped into four main temporal phases. These are, in chronological order, the gap-1 or $\mathbf{G_1}$ phase, the synthesis or **S** phase, the gap-2 or $\mathbf{G_2}$ phase and mitosis or **M** phase. While the S and M phase are the main functional units of the cell cycle in which DNA replication and nuclear division are accomplished, the intercalated gap phases represent intermediate stations during which cell fate decisions are taken and where checkpoints are known to control proper progression. Importantly proliferating cells always progress through the cell cycle phases in the same order [50].

- The **S** phase is the phase during which the genome of the cell is duplicated. In brief, DNA synthesis, initiated at the replication foci, progresses at the so called replication forks. In this micro-environment, an enzyme named helicase unwinds the double stranded DNA, which leads to the exposure of both DNA strains. This allows a further family of enzymes, termed DNA polymerases, to add to the free strains the complementary nucleotides one at a time [51]. The S phase terminates when the whole genome has been copied.

- The $\mathbf{G_2}$ phase then serves as a holding time in which proteins are accumulated that are needed for mitosis. In addition, the time between DNA synthesis and mitosis allows cells to ensure that the genome has been replicated properly [6].

- The **M** phase, even though relatively short if compared to the whole cell cycle time, is marked by dramatic events necessary for successful nuclear division. Several distinct processes, namely chromatin condensation, nuclear envelope breakdown and chromosomal DNA segregation, follow each other in rapid succession [52]. After restoration of both nuclei, cytokinesis, which culminates into the physical partition of the parental cell into two separate daughter cells, is initiated and can last in some cells for about 10 minutes [37].

- The $\mathbf{G_1}$ phase is the phase that both new-born daughter cells enter, by definition, immediately after division. In this phase, cells interrogate their environment and internal state to 'decide' whether they continue a further cycle, whether they enter a quiescent state termed $\mathbf{G_0}$ or even whether they undergo apoptosis [1]. In case that cells set into a further round of replication, DNA is primed before the onset of S phase [53].

The checkpoints

Cell cycle progression is not error-free, and an elaborate control system in form of checkpoints has evolved that ensures that errors or irregularities in the various processes are either corrected and solved or in cases in which this is not possible, that the apoptotic pathways are activated. Possible errors and problems throughout the cell cycle include damaged or wrongly replicated DNA, misaligned chromosomes, small cell size, confluence, and insufficient nutrients. In addition to their role as guardians of the

cell cycle, checkpoints also help to coordinate physically independent processes like DNA replication, growth and organelle duplication.

- The **G_1 checkpoint**, also known as the restriction point in mammalian cells and START in yeast cells, controls entry of cells into S phase. It integrates both positive and negative signals before 'green light' is given for initiation of DNA synthesis [1].
- The **G_2 checkpoint** regulates progression into mitosis by requiring that all DNA has been properly duplicated. If double-stranded DNA breaks or damage arising during replication is sensed at this point, cell cycle progression is halted, and DNA repair mechanisms are activated [6].
- The **mitotic-spindle checkpoint** ensures symmetric segregation of genomic information during cell division. It blocks the onset of sister chromatid separation until all chromatid pairs are aligned on the mitotic spindle [7].

The importance of the checkpoints for the present work lies in their potential impact on the average and on the variability in the duration of the cell cycle and the cell cycle phases. While it might be argued for instance that, under normal conditions, DNA synthesis should always progress at approximately the same rate, repairing damaged DNA might take variable time, depending e.g., on the severity or nature of the damage.

The cyclins and cyclin-dependent kinases

So far our overview on the characteristics of the cell cycle was restricted to a cellular descriptive level, which is most relevant for this work. However even a very basic introduction into the cell cycle would not be complete if cyclins and cyclin-dependent kinases would not have been mentioned. A fascinating wealth of information is available describing the 'engine' that drives the cells through each phase.
The first cyclin-dependent kinase (Cdk), called Cdk1, was discovered in genetic screens for yeast mutants with irregularities in cell division [54]. Cyclins, on their turn, had been described a few years earlier as a set of proteins that were synthesized and degraded in a cyclical fashion during cell cycle progression in sea urchin eggs [55]. Both findings could ultimately be understood as forming part of a unique and universal cell cycle control mechanism, when in 1989, Nurse et al. (Nobel laureate in 2001) reported the association of Cdk1 with Cyclin A and Cyclin B in several model organisms [56]. Since then, a whole set of highly conserved cyclins and Cdks have been identified with distinct functions during the different phases of the cell cycle [57]. While the overarching action of cyclins has been found in activating Cdks, the latter in general, upon activation, mediate processes crucial for e.g, DNA synthesis, cell growth and cytokinesis [58].
The role of the different cyclins and corresponding Cdks is to regulate cell division in a way such that cells initiate and progress through the cell cycle in an orderly fashion. This is achieved, as has been proposed based on theoretical and experimental work, by an extended network of protein interactions involving negative, positive and time-delayed negative feedback loops [47]. While the main design principles of this network are relatively well understood and have been translated into mathematical models of cell cycle regulation, the details of the potentially very large number of interactions remains an active field of research.

1.2 Measuring cell cycle parameters

There exist a plethora of experimental methods to interrogate a proliferating cell population. The method used in a given study depends on the research question at hand and sometimes also on the field. Often qualitative questions are addressed like: Are cells in a certain tissue proliferating? Are cells dividing more or less vigorously given altered experimental conditions? Are cells blocked in a certain stage of the cell cycle. More quantitative questions encompass: How many new cells are born

per hour in a given cell population? How long does it take an average cell to progress from entry into G_1 until cytokinesis? In the last 40 years, with the advent of molecular biology, asking 'Which genes are expressed in which phase?' has also become more and more important.
In this section the most common experimental methods are reviewed and their applicability, strengths and drawbacks are highlighted.

- **Light microscopy** of living cells represents probably the first and simplest method that allows to observe individual cells in cell cycle. Due to the fact that cells do condensate their chromatin during mitosis makes it possible to discern this stage relatively easily from cells in the other phases [59]. Furthermore cytokinesis can also directly be followed in some cell lines as a rounding up and a subsequent division event. Some protocols still in use today rely on the light microscope to extract information from proliferating cell cultures in vitro. For example, a simple way to test whether a cell culture is proliferating is by counting cells in so-called W-chambers. For this, cells in a fixed volume are counted under the microscope and the total population is extrapolated from these measurements. A more elaborated experimental setup, termed long-term cinematography, popular in the sixties, allows, by taking pictures of the culture dish at relative short intervals, to measure the inter-division time distribution [12]. Today this technique has been replaced by long-term video microscopy, which is however essentially the same [42].

- **Stathmokinetic analysis** is also based on the fact that cells in mitosis can be detected relatively easily by light microscopy. Progression through mitosis is artificially perturbed by administration of a cell cycle blocker that acts in mitosis (e.g., colchicine or vincristine). By estimating the relative increase in mitotic figures over time, the cell birth rate can be estimated [60]. With an appropriate model, the division time can additionally be derived [38]. The two main disadvantages of this technique are on the one hand the tedious counting of mitotic figures and on the other hand the adverse and unpredictable effects that a cell cycle blocker may have on the cellular system under study.

- **Pulse labeling with nucleoside analogs** forms the basis for several related techniques which will be described in the following. In its most simple form, a proliferating cell population is briefly exposed to a nucleoside analog like tritiated thymidine ($[^3H]$-Tdr) or the nowadays more commonly used bromodeoxyuridine (BrdU). Only those cells that are actively synthesizing DNA during the pulse incorporate the label into their newly formed DNA strands. If this population is fixed immediately after the pulse and analyzed by autoradiography in the case of radioactive thymidine, or alternatively by fluorescence-activated cell sorting (FACS) or immunohistology if BrdU was used, the fraction of labeled cells provides a good proxy for the fraction of cells in S phase. If, instead, the labeled cells are not fixed, but held under conditions in which they continue to divide (pulse-chase method), kinetics of labeled and unlabeled cells can be used to extract further properties:

  - The **grain count diminution method** exploits the fact that $[^3H]$-Tdr acquired during pulse exposure is divided approximately equally between the two daughter cells after mitosis [61]. As a consequence, the mean autoradiographic grain count of labeled cells is reduced by half after each cell division. Therefore the half-life of the initial label intensity estimated from serial samples serves as a measure for the time labeled cells take to complete the S, G_2 and M phase.

  - The **fraction of labeled mitoses method** combines the tritiated nucleoside pulse labeling technique with the possibility to measure cells that are both labeled and in mitosis [38]. The resulting kinetics are typically analyzed using computer programs to infer the duration of the cell cycle phases. The last two methods are nowadays rarely used due to the complications associated with handling radioactive material. Furthermore, the counting of

mitotic figures is time consuming, requires relative high resolution and has therefore been replaced by high-throughput methods that are based on flow cytometry.

- The **DNA-BrdU pulse-chase labeling method,** developed by Dolbeare and co-workers in 1982 [62], stains the cells additionally to BrdU with DNA specific labels like PI or DAPI. Based on bi-variate flow cytometric analysis of labeled cell populations, the synthesis time and the potential doubling time can be estimated from a single sample using for example the relative movement method [14]. As for the fraction of labeled mitosis method, mathematical models are necessary to extract cell cycle parameters from the data.

- **Long term labeling and delabeling with nucleoside analogs**, in contrast to pulse labeling methods, employs continuous labeling with nucleoside analogs, or the less toxic heavy water which gets incorporated into cells that are synthesizing DNA [43, 63, 64]. Due to the prolonged labeling over several weeks, all cells that enter the cell cycle during this period will become labeled. This method is especially useful if entry into the cell cycle is rare and many cells in the population of interest remain in a quiescent state. After the labeling phase, when a sufficiently large fraction of cells has become labeled, BrdU or deuterium administration is discontinued. The subsequent decrease in the number of labeled cells, or the decrease in the percentage of deuteriated DNA, contains information about the life-time of the labeled cells. The theoretical treatment of the labeling and delabeling curves is discussed in more detail in [65].

- **Carboxyfluorescein succinimidyl ester (CFSE) dilution** is a relatively new method that has been exploited extensively in recent years, especially in immunology [39,66], but also in other fields [67]. The cell's intracellular matrix, if exposed long enough, binds covalently and relatively uniformly to CFSE. Subsequently, each time a labeled cells divides, the amount of CFSE bound to its intracellular matrix is halved. This allows to track cell division over up to 8 generations. Usually CFSE, in contrast to BrdU, is not administrated directly to animals but is used in cell transfer experiments.

- **Gene and RNA expression assays** can be used to measure the concentrations of cell cycle related proteins and mRNA in proliferating cells. While gel-electrophoresis (northern (RNA) and western (protein) blotting) and PCR represented the methods of choice over many years, micro arrays [68–70] and deep sequencing of mRNA [71] become more and more popular nowadays. Due to the low concentrations, it is usually unavoidable to synchronize cells for bulk measurements. Nevertheless, a large part of what we know about the molecular biology of the cell cycle has been derived in studies relying on this type of experiments.

- **Fluorescence microscopy** utilizes fluorescent dies complexed to antibodies which bind surface proteins of living cells to track cycling and non-cycling cells *in vitro*. Alternatively, fluorescent reporter genes that are specifically expressed during certain cell cycle phase have been used to image in real-time cell cycle progression in genetically modified mice [28,72]. With the advent of two-photon microscopy it has also become feasible to image dividing cells in tissues *in vivo* [37]. Unfortunately the typical imaging window is currently limited to 1-3 hours, which makes it impossible to observe a single cell *in vivo* over a whole cell cycle. Furthermore if cells show a migrating phenotype, tracking of a single cell over longer periods of time becomes, due to the small volume that is typically imaged, increasingly unlikely.

1.3 Mathematical models of the cell cycle

Most if not all theoretical models of the cell cycle can be interpreted as compartmental models. In general they distinguish themselves in the number of compartments, in the allowed transitions between

the compartments, in the completion time distribution for each compartment and finally in the way they are defined, analyzed or implemented.
In the following, we will restrict our attention to cell cycle models at the cell population level, with the most important variables being the number of cells and their respective state (e.g., phase, generation). Models describing for instance explicitly gene regulatory networks or DNA synthesis are out of the scope of the present work.
Probably the most trivial cell cycle model at the cell population level is given by a single compartment model assuming an exponentially distributed completion time[1] distribution with mean α and zero cell loss. This allows for a straightforward prediction of the number of cells over time, i.e.,

$$n(t) = n_0 \, e^{\frac{1}{\alpha} t}. \tag{1.1}$$

The latter is the solution of the following ordinary differential equation

$$\dot{n}(t) = \frac{1}{\alpha} \, n(t); \quad n(0) = n_0,$$

with $\frac{1}{\alpha}$ being the so-called Malthusian constant.
As a simple and possibly more realistic alternative, the completion or the division time for cells may be assumed as fixed instead of exponentially distributed, for example at a deterministic value of β. Then the number of non-synchronously dividing cells evolves over time according to

$$n(t) = n_0 \, 2^{\frac{1}{\beta} t} = n_0 \, e^{ln(2) \frac{1}{\beta} t}. \tag{1.2}$$

Notice that the cell populations are growing, under the two models, with an exponentially increasing rate. However, even if the average division time for both would be equal, i.e., $\alpha = \beta$, the exponential model would grow with a $1/\ln(2) = 1.44$ steeper log-growth curve.
Population kinetics predicted by Eq. 1.1 or Eq. 1.2 are well supported by experimental studies in which exponentially growing cell populations are routinely observed *in vitro* under ideal conditions of virtually unlimited space and nutrients. Such necessarily temporally restricted behavior is then termed either exponential, free, balanced or log phase growth. Typically confluence, i.e., high cell densities, limits free growth at a certain stage because it induces most cell types to undergo apoptosis or to exit the cell cycle and enter a quiescent state. Simple models accounting for this self-limiting behavior are common and have been studied for a long time [73].
The exponential and the deterministic model, even though describing accurately the kinetics of freely proliferating cell populations, are less successful in reproducing frequency distributions of intermitotic times obtained by time-lapse cinematography of exponentially growing cultures [12]. Therefore Smith *et* al. proposed a slightly more complex 2-compartment cell cycle model with compartments termed A and B. While the completion time in the A compartment was assumed exponentially distributed, the completion time in the B state was fixed. Even though later studies [42, 74–76] have shown that the model predictions do not exactly match experimental data, its simplicity and mathematical tractability makes the Smith-Martin model even today one of the most popular theoretical models in the field [77, 78].
Since then, prompted by new experimental techniques, more elaborate models have been developed. Two examples where experimental techniques lead to model refinements are the previously described DNA-BrdU pulse-chase and the CFSE dilution method. In the first case, the fact that cell cycle phases could be measured and cells in each phase followed over time, 'asked' for the inclusion of phase specific compartments into kinetic cell cycle models [18–20, 24]. In the case of CFSE dilution, the generation

[1] Throughout this work, we will use 'completion time' to denote the time it takes a cell between entry into and exit out of a given compartment. Common synonyms found in the theoretical literature are waiting time, dwelling time and passage time, among others.

number became the most important 'new variable' that could be extracted from experimental data. Generation structure, activation times and generation dependent cell death were accounted for and subsequently estimated mostly in the context of lymphocyte activation and proliferation [39,65,66,79].

Not so much inspired by technical innovations but rather motivated by the relevant clinical problem of optimizing chemotherapy and radiotherapy, a large body of work has been dedicated to model cell cycle in cancer cells [19, 80–83]. Due to a wide-spread use of cell cycle specific anti-cancer drugs, phase resolved models are quite common in this field. Optimal dose schedule design for therapeutics have been derived, which aim at maximally harming cancerous while sparing healthy cells [84]. Unfortunately these models have not yet found their wide application in the clinics.

Completion time distributions other than the most commonly assumed exponentially distributed or fixed have also been studied. Gaussian [85], Log-normal [19,23] and Gamma distributed [39] completion times represent the most frequent alternatives. These distributions, in contrast to exponentially distributed completion times, do not allow for a mathematical treatment using ODE models. Therefore new methods were developed and existing methods based on general probability theory have been adapted to understand these models. One popular approach are age-structured models [76, 86]. They are usually defined as partial differential equations, for example in the following way

$$\frac{\partial n(t,a)}{\partial t} + \frac{\partial n(t,a)}{\partial \tau} = D(a)n(t,a),$$

where $n(t,a)$ is the cell number density of cells with maturity age a at time t, and $D(a)$ is an time-independent transition matrix. Given an initial distribution $n(0,a)$ the evolution of $n(t,a)$ can be derived using techniques typically used to solve partial differential equations. If analytical solutions are too difficult or impossible to obtain, which is the rule rather than the exception, numerical algorithms, like cellular automaton models [18, 21, 87] or the method of successive generations [19] can be implemented in order to compare predictions with experimental results. Another approach employs the theory of branching processes in order to cope with non-exponential completions time distributions [86, 88]. Although solvable only for highly simplified models, this powerful technique allows, in addition to the average kinetics, to predict fluctuations in the abundances of cells in the various compartments.

A further class of cell cycle models describes kinetics, specific to cell populations under homeostatic control. To give an example, the following ODE model was developed by Mohri *et* al. [43] to analyze long-term BrdU labeling experiments of CD-4 and CD-8 T cells in SIV-infected rhesus macaques. For the labeling period, unlabeled (U) and labeled (L) cells were modeled by

$$\dot{U} = s_U - pU - dU \qquad \dot{L} = s_L + 2pU + pL - dL$$

and for the delabeling period

$$\dot{U} = s_U^{'} + pU - dU \qquad \dot{L} = s_L^{'} + pL - dL.$$

Here s_U, s_L, $s_U^{'}$, $s_L^{'}$ represent source terms (e.g., the thymus), p is the proliferation rate and d is the death rate. Analytical solutions have been derived for this system, which were then compared to experimental data. Subsequently, several modification of this model have been proposed to account for different experimental conditions or assumptions, leading to interesting insights into the dynamics of T lymphocytes during SIV and HIV infection [63, 89].

1.4 Results

1.4.1 A phase resolved stochastic cell cycle model

Model definition

We describe the eukaryotic cell cycle as an orderly sequence of three phases, distinguished by cellular DNA content, conventionally termed G_1, S and G_2M[2] (see Section 1.1.1). Proliferating cells are supposed to proceed, under this minimalist view, from one phase to another in a fixed order, until reaching the end of G_2M phase. At this point, they undergo cytokinesis which generates, starting with a single parental cell, two genetically identical daughter cells. These are by definition in G_1 phase as soon as cell division is completed (Fig. 1.1 B). In addition, we assume that the completion time in any given phase (i.e., the time lapse between the entry and exit in the given phase) is a random variable τ, which is distributed according to a shifted-exponential density function (Fig. 1.1 A):

$$f_\tau(\tau) = \frac{1}{\alpha} e^{-\frac{1}{\alpha}(\tau-\beta)} H(\tau-\beta), \tag{1.3}$$

where α is the reciprocal of the rate of the exponential, β is the fixed delay and H denotes the heavyside step function whose value is zero for negative argument, i.e., for $t<\beta$, and one for positive argument. Notice that with a slight abuse of notation we denote here the random variable and the value it assumes by the same symbol τ. The delay β in Eq. 1.3 'ensures' that a cell that enters the specific phase will remain therein for at least β time units (e.g., hours) before proceeding to the next phase. Besides this fixed minimal time β, additional less predictable effects that affect the completion of the processes associated to a phase are assumed to be exponentially distributed with both mean and standard deviation given by α. The phase specific average completion time, denoted in the following by $\bar{\tau}$, is then $\alpha+\beta$ with standard deviation α and coefficient of variation $\alpha/\bar{\tau}$. The Laplace transform of Eq.1.3 is given by

$$\mathcal{L}_\omega(f_\tau(\tau)) = \frac{e^{-\beta\omega}}{1+\alpha\,\omega}. \tag{1.4}$$

where ω is the transformed variable corresponding to the time lapse τ. The temporal organization of the cell cycle is defined by the vector of phase-specific completion times $\boldsymbol{\tau} = \{\tau_{G_1}, \tau_S, \tau_{G_2M}\}$, which in turn depend on the parameter vectors $\boldsymbol{\alpha} = \{\alpha_{G_1}, \alpha_S, \alpha_{G_2M}\}$ and $\boldsymbol{\beta} = \{\beta_{G_1}, \beta_S, \beta_{G_2M}\}$. For the average phase durations we have $\bar{\boldsymbol{\tau}} = \{\alpha_{G_1}+\beta_{G_1}, \alpha_S+\beta_S, \alpha_{G_2M}+\beta_{G_2M}\}$.
The cell cycle length, understood as the time lapse between the entry into G_1 until exit out of G_2M, is also a random variable, and will be denoted herein by $T = \tau_{G_1} + \tau_S + \tau_{G_2M}$. Its probability density function (PDF) is the convolution of the three underlying shifted exponential distributions and corresponds to the so-called shifted hypoexponential distribution. Explicit expressions can be computed using the inverse Laplace transform $\mathcal{L}^{-1}$ of the product of the Laplace transforms of the three densities given by Eq. 1.4, i.e.,

$$f_T(T) = \mathcal{L}_T^{-1}\Big(\prod_{i=G_1}^{G_2M} \frac{e^{-\beta_i\omega}}{1+\alpha_i\,\omega}\Big) \quad \text{with} \quad i \in \{G_1, S, G_2M\}. \tag{1.5}$$

In case that all entries in $\boldsymbol{\alpha}$ are distinct, we get

$$f_T(T) = \sum_i \left(\frac{\alpha_i\, e^{-\frac{(T-B)}{\alpha_i}}}{\prod_{j,\, i\neq j}(\alpha_i-\alpha_j)} \right) H(T-B), \tag{1.6}$$

[2]Based on DNA content, the G_2 and M phase are indistinguishable. Therefore we combine in our model both phases into a single compartment.

in which the indices i and j iterate over the three phases and B is the sum over the elements in $\boldsymbol{\beta}$. The corresponding cumulative distribution function (CDF) is computed as

$$F_T(T) = \int_B^T f_T(x)dx = \sum_i \left(\frac{\alpha_i^2 \left(1 - e^{-\frac{(T-B)}{\alpha_i}}\right)}{\prod_{j,\, i \neq j}(\alpha_i - \alpha_j)} \right) H(T - B). \tag{1.7}$$

Notice that the lower integration limit in the integral above is B and not $-\infty$, because, according to Eq. 1.6, $f_T(T) = 0$ for $T < B$.
The shape of the phase specific completion time distribution f_τ , shown in Fig. 1.1 A and defined by Eq. 1.3, visualizes the fact that the probability for a cell to complete a given phase in less than β time units is zero. Also notice that each phase can have distinct parameter values α and β for its completion time distribution.

A preliminary experimental validation

The theoretical predictions obtained in the previous section can be utilized to perform a preliminary validation of the model by comparing Eq. 1.6 and an expression involving Eq. 1.7 to appropriate experimental data.
Careful measurements of proliferating cell lines *in vitro* have been collected by e.g., Smith *et* al. ([12], 1965) and more recently by Hawkins *et* al. ([42], 2009) using time-lapse cinematography and long-term video tracking respectively. The data points reproduced in Fig. 1.2 A represent, for several distinct cell lines, the percentages of undivided cells, given birth at time zero (see [12] for further details). This data served initially to motivate the now famous Smith-Martin model (see Section **??**), which posits a shifted-exponential distribution, as defined in Eq. 1.3, not for each phase, but for the total cell cycle length. The histogram in Fig. 1.2 B, on the other hand, shows the division time PDF obtained by continuous observation of CpG-stimulated proliferating B cells [42].
The kinetics in Fig.1.2 A are for certain parameter values well approximated by $100 \times (1 - F_T)$. The latter expression corresponds to the predicted percentage of undivided cells with time, given birth at time zero. Similarly the empirical division time histogram in Fig. 1.2 B is closely reproduced by the shifted hypoexponential distribution defined in Eq. 1.6. Obviously this cannot be taken as a proof that the proposed model is 'correct', a point which is underpinned by the fact that alternative and even simpler models like the shifted log-normal and the shifted gamma distribution [42] are able to reproduce the distribution equally well (not shown). While the two latter distributions depend on three parameters each, the shifted hypoexponential distribution depends, in our case, on six parameters. Not surprisingly, the parameters remain given this kind of data largely undetermined. To test whether the model assumptions concerning the phase completion times are reasonable, further theoretical predictions and experimental data are therefore needed.

1.4.2 Parameter identification under balanced growth

Balanced growth

A proliferating cell population that obeys the probability model specified in the previous section can be represented by a non-Markov multidimensional random process, whose evolution depends on its history. There exist an infinite number of possible histories of the population size $N(t)$. We focus here on a specific important subset, namely those under balanced growth. The latter are implicitly defined by an in average exponentially growing cell population $E[N(t)] \stackrel{\text{def}}{=} E[N_{\mathrm{G}_1}(t) + N_{\mathrm{S}}(t) + N_{\mathrm{G}_2\mathrm{M}}(t)] \propto e^{\mu t}$ with growth rate μ and constant average fractions of cells in each phase $\boldsymbol{n} = \{n_{\mathrm{G}_1}, n_{\mathrm{S}}, n_{\mathrm{G}_2\mathrm{M}}\}$, where e.g., $n_{\mathrm{G}_1} \stackrel{\text{def}}{=} E[N_{\mathrm{G}_1}(t)/(N_{\mathrm{G}_1}(t) + N_{\mathrm{S}}(t) + N_{\mathrm{G}_2\mathrm{M}}(t))]$. The expectation operator $E[\cdot]$ is defined over all

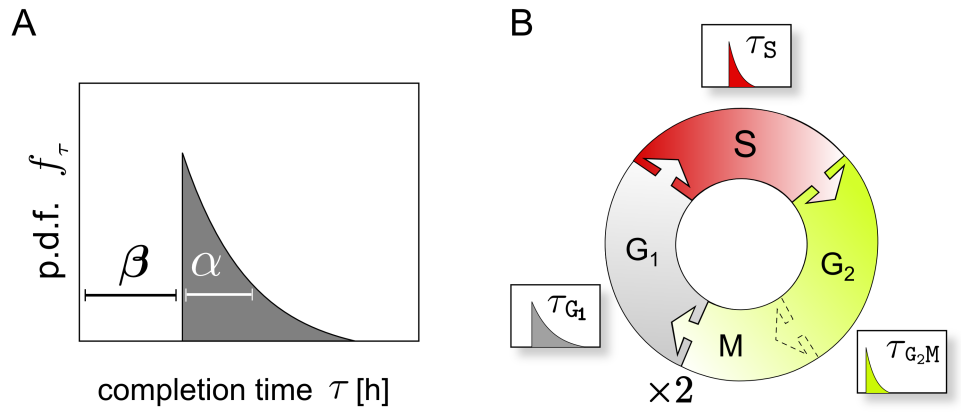

Figure 1.1: Stochastic cell cycle model. (A) Shifted-exponential completion time distribution density f_τ with parameters α and β. (B) Scheme of the proposed cell cycle model with three phases G_1, S and G_2M. The dashed border between the G_2- and the M-phase indicates that the G_2 and M-phase are combined into a single phase. The random time τ a cell needs to complete the processes associated with each phase follows a shifted exponential distribution with phase-specific parameters α and β.

possible realizations of the process. It follows from the definition of balanced growth that the average history of the population size $N(t)$ is also proportional to $e^{\mu t}$.

We now derive explicit expressions for n_{G_1}, n_S and n_{G_2M} and a transcendental equation that defines μ, the balanced growth rate. A first step in obtaining the constant frequencies of cells in each of the phases consist in computing the ratio between the cells that complete a given phase and the total number of cells inside the same phase at time t. This vector quantity, denoted here by $\boldsymbol{\gamma}$, represents the asymptotic efflux rate constant, which will be useful, as we will see, to construct a transition probability matrix $\mathcal{Q}$. The latter will enable us to employ methods from linear algebra to solve the steady state condition.

Suppose for example that a cohort of cells entered a given phase i at time t_{in}. Then a proportion $f_{\tau_i}(t_{out} - t_{in})$ of these cells will leave this phase at time t_{out}. Similarly if a cohort of cells entered a phase at time t_{in} then a proportion $1 - \int_0^{t-t_{in}} f_{\tau_i}(x)dx \overset{\text{def}}{=} R_{\tau_i}(t - t_{in})$ will remain in this phase until time t, where R denotes the so-called reliability function.

Recalling that the influx of cells into a given phase is proportional to $e^{\mu t}$, integrating over all past entries and finally taking the ratio we obtain for $i \in \{G_1, S, G_2M\}$

$$\gamma_i = \frac{\int_{-\infty}^{t} e^{\mu x} f_{\tau_i}(t-x)dx}{\int_{-\infty}^{t} e^{\mu x} R_{\tau_i}(t-x)dx} = \frac{\mu \mathcal{L}_\mu\{f_{\tau_i}\}}{1 - \mathcal{L}_\mu\{f_{\tau_i}\}} = \frac{\mu}{(\alpha_i \mu + 1)\, e^{\beta_i \mu} - 1}, \tag{1.8}$$

where we used for the second equality Eq. 1.36 and Eq 1.37. For a phase without a delay, i.e., $\beta_i = 0$, this expression simplifies to the more familiar mass action principle, where the transition probability is directly proportional to the decay rate α_i^{-1}. Assuming that cells are immortal and recalling that division occurs as cells proceed from G_2M to G_1, we build up the transition probability matrix as follows

$$\mathcal{Q} = \begin{bmatrix} -\gamma_{G_1} & 0 & 2\,\gamma_{G_2M} \\ \gamma_{G_1} & -\gamma_S & 0 \\ 0 & \gamma_S & -\gamma_{G_2M} \end{bmatrix} \tag{1.9}$$

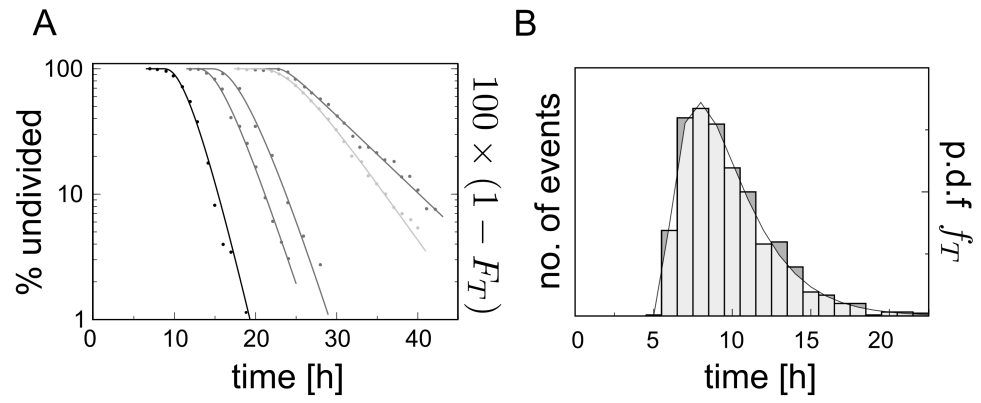

Figure 1.2: Preliminary experimental validation. (A) Best fit of the model predictions, i.e., $100 \times (1 - F_T)$, to the percentage of undivided cells after birth obtained by time lapse cinematography [12] of slow and fast dividing cell lines. (B) Best fit of f_T defined by Eq. 1.6 (solid line) to inter-mitotic time distribution density measured by long-term video tracking of *in vitro* proliferating B-cells [42]. The data in C and D were extracted from the graphs in the original publications as accurately as possible.

The balanced growth condition can now be formulated in matrix form

$$\mathcal{Q} \begin{bmatrix} n_{\mathrm{G}_1} \\ n_{\mathrm{S}} \\ n_{\mathrm{G}_2\mathrm{M}} \end{bmatrix} = \mu \begin{bmatrix} n_{\mathrm{G}_1} \\ n_{\mathrm{S}} \\ n_{\mathrm{G}_2\mathrm{M}} \end{bmatrix} \tag{1.10}$$

where μ is an eigenvalue of $\mathcal{Q}$ and $\boldsymbol{n} = \{n_{\mathrm{G}_1}, n_{\mathrm{S}}, n_{\mathrm{G}_2\mathrm{M}}\}$ is the corresponding eigenvector. It can be shown that there exist a single dominating real positive eigenvalue for $\mathcal{Q}$ (see Material and Methods, Stability Analysis) whose associated normalized eigenvector is

$$\begin{bmatrix} n_{\mathrm{G}_1} \\ n_{\mathrm{S}} \\ n_{\mathrm{G}_2\mathrm{M}} \end{bmatrix} = \begin{bmatrix} 2 - 2\,\exp(-\beta_{\mathrm{G}_1}\mu) \times (\alpha_{\mathrm{G}_1}\mu + 1)^{-1} \\ 2 - n_{\mathrm{G}_1} - \exp(\beta_{\mathrm{S}}\mu) \times (\alpha_{\mathrm{S}}\mu + 1) \\ -1 + \exp(\beta_{\mathrm{G}_2\mathrm{M}}\mu) \times (\alpha_{\mathrm{G}_2\mathrm{M}}\mu + 1) \end{bmatrix} \tag{1.11}$$

The uniqueness and existence of a dominating positive real root ultimately motivates our focus on balanced exponential growth, as any immortal proliferating cell population with sufficient nutrients and space will eventually enter this stationary phase. The time it takes, either starting with a single cell or a synchronized cell cohort to enter this state depends on the cell cycle parameters. This aspect, although important, will not be further investigated here. The exponential growth rate μ is the unique real positive root of the characteristic equation $\det(\mathcal{Q} - \mu\mathbb{I}) = 0$ which writes as

$$\frac{\mu^3(2 - \prod_i e^{\beta_i\mu}(1 + \alpha_i\mu))}{\prod_i (e^{\beta_i\mu}(1 + \alpha_i\mu) - 1)} = 0. \tag{1.12}$$

It is easy to see that the denominator in Eq. 1.12 is always positive. To determine a non-trivial μ it therefore remains to solve the transcendental equation in the nominator

$$2 - \prod_i e^{\beta_i\mu}(1 + \alpha_i\mu) = 0, \tag{1.13}$$

which does not allow for any analytical solution. Numerical solutions can be computed using e.g., the Newton-Raphson root finding algorithm, with fast convergence if the initial value is set to $\mu_0 = \log(2)/\overline{T}$,

where $\overline{T}$ is the average cell cycle length, i.e., the sum of the elements in $\bar{\boldsymbol{\tau}}$. This first guess is a naive estimate for μ assuming that cells divide according to a deterministic division time identical to the average of the hypoexponential density defined in Eq. 1.5.

Learning from balanced growth

The predictions for the fraction of cells in each of the phases can be compared to frequencies extracted experimentally from bivariate analysis of cell populations transiently exposed to nucleotide analogs (e.g., BrdU) and subsequently examined both for the intensities of the signals due to incorporated nucleotide analog and total DNA content (see [14] and Section 1.2). The question that we want to address in this section is: What can potentially be learned about the parameters of the model, given this type of experimental data? By definition the measured frequencies will sum to one, and therefore we have for three populations effectively only two equations but six model parameters. This makes it impossible to identify all the parameter values, irrespective of the number of samples we take. It is however possible to derive analytical expressions for the upper and lower bounds for both the parameters and the average completion time of each phase.

Solving Eq. 1.11 using measured frequencies denoted by $\widetilde{\boldsymbol{n}} = \{\widetilde{n}^{\,\mathrm{G_1}}, \widetilde{n}^{\,\mathrm{S}}, \widetilde{n}^{\,\mathrm{G_2M}}\}$, we get for the reciprocal of the rate parameter vector for phase $i \in \{G_1, S, G_2M\}$

$$\alpha_i = (\kappa_i e^{-\beta_i \mu} - 1)/\mu, \tag{1.14}$$

$$\text{with} \quad \boldsymbol{\kappa} = \left\{\kappa_{\mathrm{G_1}}, \kappa_{\mathrm{S}}, \kappa_{\mathrm{G_2M}}\right\} = \left\{\frac{2}{2-\widetilde{n}^{\mathrm{G_1}}}\,,\, \frac{2-\widetilde{n}^{\mathrm{G_1}}}{1+\widetilde{n}^{\mathrm{G_2M}}}\,,\, 1+\widetilde{n}^{\mathrm{G_2M}}\right\}. \tag{1.15}$$

Both α_i and the delay parameters β_i are by assumption greater or equal zero. These conditions propagate into Eq. 1.14 which allows us to specify boundaries for α_i and β_i. First notice that α_i is a monotonically decreasing function in β_i with its maximum $(\kappa_i - 1)/\mu$ at $\beta_i = 0$ and a zero crossing at $\beta_i = \ln(\kappa_i)/\mu$. The maximum and the root represent the upper bounds for α_i and β_i respectively, while the lower bounds are zero for both. We thus have

$$\alpha_i \in [0, (\kappa_i - 1)/\mu] \quad \text{and} \quad \beta_i \in [0, \ln(\kappa_i)/\mu]\,. \tag{1.16}$$

The mean phase-specific completion time, $\bar{\tau}_i$, the sum of the reciprocal rate vector α_i and the delay vector β_i, is also bounded, with an interval given by

$$(\alpha_i + \beta_i) \in [\ln(\kappa_i)/\mu, (\kappa_i - 1)/\mu]\,. \tag{1.17}$$

This result is derived from the fact that $(\alpha_i + \beta_i)$ is concave having its unique minimum at $\beta_i = ln(\kappa_i)/\mu$, which follows from setting the derivative $\partial_\beta(\alpha_i + \beta_i) = 1 - \kappa_i e^{-\beta_i \mu}$ to zero. This implies that $(\alpha_i + \beta_i)$ is a monotonically decreasing function in the interval $\beta_i \in [0, \ln(\kappa_i)/\mu]$ with the corresponding extrema specified above. It is important to note that intervals defined by Eqs 1.14-1.17 depend on the average growth rate μ which is in general not known. Formally, if one specific pair of parameter vectors α_i and β_i explains the measured frequencies with growth rate μ, the scaled parameter vectors $c\,\alpha_i$ and $c\,\beta_i$ mimic equally well the same data for arbitrary positive c, however with a reduced growth rate μ/c. This can be easily verified by substituting these expressions in Eq. 1.11 and Eq. 1.13. The consequence is that μ remains undefined. However for the relative average time a cells spends e.g., in $\mathrm{G_1}$ phase $(\alpha_{\mathrm{G_1}} + \beta_{\mathrm{G_1}})/\overline{T}$ the growth rate cancels out.

Using the fact that $\kappa_i \in\,]1, 2[$ and the appropriate series expansion for the natural logarithm, the

width of the intervals bounding α_i, β_i and $(\alpha_i + \beta_i)$ can be rewritten as:

$$\begin{aligned} w_{\alpha_i} &= (\kappa_i - 1)/\mu = \sum_{i=1}^{1} \frac{(-1)^{i+1}}{i} (\kappa_i - 1)^i/\mu, \\ w_{\beta_i} &= \sum_{i=1}^{\infty} \frac{(-1)^{i+1}}{i} (\kappa_i - 1)^i/\mu, \\ w_{\alpha_{i+\beta_i}} &= \sum_{i=2}^{\infty} \frac{(-1)^{i}}{i} (\kappa_i - 1)^i/\mu. \end{aligned} \tag{1.18}$$

From this it can be proven that $w_{\alpha_i} > w_{\beta_i} > w_{\alpha_i+\beta_i}$. This shows that by using measurements of the phase-specific stationary cell frequencies to infer the phase-specific completion times τ results in estimates of the mean value $\alpha_i + \beta_i$ that are more precise than the estimates of the variability, i.e. of the standard deviation α_i. Notice that the width of the intervals can be interpreted as a naive lower bound for the uncertainty about the respective parameter values. For the two data sets that we will analyze later, we compute intervals for the phase-specific standard deviations w_{α_i} which are on average ≈ 10 times wider than the intervals for the expected phase-specific completion times $w_{\alpha_i+\beta_i}$.

1.4.3 Parameter identification under unbalanced growth

Unbalanced growth

Balanced growth analysis does not allow to distinguish between fixed and purely exponentially distributed completion times even when μ is known. This follows from Eq. 1.16, because, for any phase i, possible values for the standard deviation α_i include 0 (fixed) and $(\kappa_i - 1)/\mu$, which due to Eq. 1.17 requires the delay β_i to become 0 (purely exponential).

A possible avenue that overcomes this limitation is perturbation of the stationary phase. To illustrate this we consider a simple *Gedankenexperiment,* that consists in separating physically cells from a population under balanced growth, at a given time, according to a specific phase, say ϕ, which could be either G_1, S or G_2M. This unbalanced or synchronized cell population is, under our model assumptions, predicted to return to balanced growth, i.e., to the restoration of exponential growth kinetics and to the stationary proportions of cells in each phase. It turns out that measuring after the partition the transient kinetics of this population yields information that potentially allows, in contrast to balanced growth analysis, to distinguish between a fixed and a purely exponentially distributed phase completion time. More specifically a mathematical proof will show that taking samples at three well placed support points permits under ideal conditions accurate estimation of the average and the variability in the time required to complete the phase ϕ.

The initial average fraction of cells in ϕ is specified by the corresponding entry in the eigenvector defined by Eq. 1.11. To predict when and at which rate the isolated cells will have completed ϕ we need to specify their histories first. For the time before the partition the average influx into ϕ is proportional to $e^{\mu t}$ under balanced growth. After the partition, i.e., for $t > t_0$, the influx vanishes, but only for those cells which completed zero phases since t_0. In the following we will denote the number of phases a cell cohort completed since t_0 by p. Cells which completed at least one phase (i.e., $p > 0$) since t_0, are treated, as we shall see, as distinguishable subpopulations, whose kinetics will be considered as soon as the fate of the cells with $p = 0$ has been determined.

In close analogy to expression Eq. 1.8 we compute the time-dependent flux density for cells with $p = 0$ into the compartment of cells with $p = 1$ as

$$\boldsymbol{\gamma}^{\phi}_{0\to 1}(t) = \frac{\int_{-\infty}^{t_0} e^{\mu x} f_{\tau_\phi}(t - x)dx}{\int_{-\infty}^{t_0} e^{\mu x} R_{\tau_\phi}(t_0 - x)dx}. \tag{1.19}$$

Notice that the upper integration limit is t_0 and not t, because the influx vanishes after t_0. On the left-hand side, the arrow from 0 to 1 indicates that this expression describes the flux density for cells from zero ($p = 0$) to one ($p = 1$) phases completed after partitioning, thus corresponding to the frequency of cells (relative to the initial number of cells in ϕ) which complete ϕ at time t.

After evaluating both integrals, Eq. 1.19 yields for $t > t_0 = 0$

$$\boldsymbol{\gamma}^{\phi}_{0\to1}(t) = \frac{\mu\mathcal{L}_{\mu}\{f_{\tau_\phi}(x+t)\}}{1-\mathcal{L}_{\mu}\{f_{\tau_\phi}\}} = \begin{cases} \dfrac{\mu\, e^{\mu t}}{(1+\alpha_\phi\mu)\, e^{\beta_\phi\mu}-1} & t < \beta_\phi \\ \\ \dfrac{\mu\, e^{\frac{1}{\alpha_\phi}(\beta_\phi+\alpha_\phi\beta_\phi\mu-t)}}{(1+\alpha_\phi\mu)\, e^{\beta_\phi\mu}-1} & t \geq \beta_\phi \end{cases}, \tag{1.20}$$

where for the first equality we used Eq.1.40 and Eq.1.41. It follows that the accumulated average cell flux that at time t has completed ϕ and progressed to the next phase is given by

$$\Gamma^{\phi}_{0\to1}(t) = \int_{t_0}^{t} \boldsymbol{\gamma}^{\phi}_{0\to1}(x)\, dx, \tag{1.21}$$

which for $t \to \infty$ approaches one, reflecting the fact that, in our model, all cells will eventually complete ϕ.

The Laplace transform of Eq.1.21 writes as

$$\mathcal{L}_{\omega}\left\{\Gamma^{\phi}_{0\to1}(t)\right\} = \frac{\mu+\alpha_\phi\omega\mu-\mu e^{\beta_\phi(\mu-\omega)}(1+\alpha_\phi\mu)}{\omega(e^{\beta_\phi\mu}(1+\alpha_\phi\mu)-1)(\omega-\mu)(1+\alpha_\phi\omega)},$$

where ω is again the transformed variable corresponding to t.

Within a cohort of cells isolated for instance in S phase, i.e., $\phi = \mathrm{S}$, the accumulated average cell flux out of the subsequent G_2M phase can then be derived recalling Eq.1.4 and using the properties of the inverse Laplace transform as

$$\Gamma^{\mathrm{S}}_{1\to2}(t) = \mathcal{L}^{-1}_{t}\left\{\mathcal{L}_{\omega}\left\{\Gamma^{\mathrm{S}}_{0\to1}(t)\right\} \times \frac{e^{-\beta_{\mathrm{G_2M}}\omega}}{1+\alpha_{\mathrm{G_2M}}\,\omega}\right\}. \tag{1.22}$$

For an arbitrary cell cohort, the accumulated average flux completing p phases and entering the $(p+1)^{th}$ phase since isolation can be written in general as

$$\Gamma^{\phi}_{p\to p+1}(t) = \mathcal{L}^{-1}_{t}\left\{\mathcal{L}_{\omega}\left\{\Gamma^{\phi}_{0\to1}(t)\right\} \times \prod_{i=1}^{p} \frac{e^{-\beta_{q(\phi,i)}\omega}}{1+\alpha_{q(\phi,i)}\omega}\right\}, \tag{1.23}$$

where $q(\phi, i)$ denotes a function with $i \in \mathbb{N}$ and $q(\phi, 0) \overset{\text{def}}{=} \phi$. For $i > 0$, $q(\phi, i)$ is recursively defined as the cell cycle phase which follows $q(\phi, i-1)$ according to the cell cycle model, e.g., S follows G_1, G_1 follows G_2M, and G_2M follows S. For instance, if $\phi = S$, then by definition $q(\mathrm{S}, 0) = \mathrm{S}$, $q(\mathrm{S}, 1) = \mathrm{G_2M}$, $q(\mathrm{S}, 2) = \mathrm{G_1}$ and $q(\mathrm{S}, 3)$ is again S. Notice that Eq. 1.22 corresponds to Eq. 1.23 for $\phi = S$ and $p = 1$.

Analytical expression for Eq. 1.23, although solved relatively easily with modern algebra software, can become quite cumbersome for values of p larger than six. In our case, deriving the expressions for p up to a value of five was sufficient to simulate the experiments.

Because we want to compare the model predictions with experimental determined cell frequencies, more interesting than the accumulated fluxes are the average relative numbers of cells inside each phase over time. These can be derived from Eqs 1.21-1.23 closely following the methodology outlined

in [39, 78]. For the fraction of cells initially in ϕ phase we have

$$n_0^{\phi}(t) = \frac{n^{\phi}}{e^{\mu t}} \times (1 - \Gamma^{\phi}_{0\to 1}(t)), \tag{1.24}$$

where the index 0 in $n_0^{\phi}(t)$ indicate that this expression describes cells which completed zero phases since t_0. The first term on the right hand side corresponds to the fraction of cells in phase ϕ at t_0 divided by $e^{\mu t}$, which accounts for the total population growth during the same interval. The second term stands for the fraction of cells that remained in phase ϕ up to time t relative to the initial number of cells in this phase. By evaluating the integral in Eq. 1.21, substituting in Eq. 1.24 and letting as before, without loss of generality, the time of partition t_0 be zero, we get for $t > 0$

$$n_0^{\phi}(t) = \frac{n^{\phi}}{e^{\mu t}} \times \begin{cases} 1 + \dfrac{1 - e^{\mu t}}{(1 + \alpha_{\phi}\mu)\, e^{\beta_{\phi}\mu} - 1} & t < \beta_{\phi} \\ \dfrac{\alpha_{\phi}\mu\, e^{\frac{1}{\alpha_{\phi}}(\beta_{\phi} + \alpha_{\phi}\beta_{\phi}\mu - t)}}{(1 + \alpha_{\phi}\mu)e^{\beta_{\phi}\mu} - 1} & t \geq \beta_{\phi} \end{cases}. \tag{1.25}$$

Expressions for cells initially in S, G_1 or G_2M phase can be obtained by substituting ϕ by the respective phase.

If there were no cell division (i.e., $\mu = 0$) we could readily obtain the average fraction of cells that completed p phases at time t as the difference between the cells that entered the p^{th} phase, i.e., $\Gamma_{p-1\to p}(t)$, and those that left it, i.e., $\Gamma_{p\to p+1}(t)$, divided by $e^{\mu t}$. To account for cell division, we need to multiply this difference by an additional term λ_p^{ϕ}, which increase by a factor 2 each time cell cohorts make a transition from $G_2M \to G_1$. This term is defined, for each case, as follows: $\lambda_p^{G_1} = 2^{\lfloor \frac{p}{3} \rfloor}$, $\lambda_p^{S} = 2^{\lfloor \frac{p+1}{3} \rfloor}$ and $\lambda_p^{G_2M} = 2^{\lfloor \frac{p+2}{3} \rfloor}$, where the brackets in the exponent represent the floor operator.

In general we get for all consecutive phases for cells initially in phase ϕ the relatively manageable expression

$$n_p^{\phi}(t) = \frac{\lambda_p^{\phi} \times n^{\phi}}{e^{\mu t}} \times (\Gamma^{\phi}_{(p-1)\to p}(t) - \Gamma^{\phi}_{p\to(p+1)}(t)). \tag{1.26}$$

As for Eq. 1.25 the resulting solutions are defined as piecewise-continuous functions in time.

Learning from unbalanced growth

In this section, we will show that data from the transient kinetics generated by our *Gedankenexperiment* allows to identify the average and the variability in the individual completion times. The proof is based on the analytical expressions derived in the previous section, and also on the assumption that the kinetics are acquired under the ideal conditions of no measurement errors. The latter condition, although obviously unrealistic, can, assuming unbiased measurement noise, always be approached in practice by increasing the number of samples collected at each support point.

As before, consider a cohort of cells isolated at $t_0 = 0$ in a specific phase ϕ. Substituting α_{ϕ} using Eq. 1.14 in the upper row of Eq. 1.25 and solving for μ one finds

$$\mu = \frac{\log(\kappa_{\phi}\widetilde{n}^{\phi}) - \log\left(\widetilde{n}^{\phi} + (\kappa_{\phi} - 1)\widetilde{n}_0^{\phi}(t_{]0,\beta_{\phi}[})\right)}{t_{]0,\beta_{\phi}[}}, \tag{1.27}$$

where $t_{]0,\beta_{\phi}[}$ denotes an arbitrary time point that lies in the interval $]0, \beta_{\phi}[$, $\widetilde{n}^{\phi} = \widetilde{n}_0^{\phi}(0)$, and $\widetilde{n}_0^{\phi}$ is the experimentally determined equivalent of Eq. 1.25. This shows that the balanced growth rate μ

is fully determined by only two support points, one immediately after the partition at $t = 0$ and a second at an arbitrary time point $t_{]0,\beta_\phi[}$. This also makes clear that placing more support points in the interval $t_{]0,\beta_\phi[}$ does, under ideal conditions, neither increase knowledge about μ nor the parameter values. Importantly the uncertainty about the phase-specific variability discussed in previous sections remains.
By replacing the same expression for α_ϕ in the second row of the left hand side of Eq. 1.25 we get

$$\widetilde{n}_0^{\phi}(t_{[\beta_\phi,\infty]}) = \frac{\widetilde{n}^{\phi} e^{\left(\frac{\kappa_\phi \mu t_{[\beta_\phi,\infty]} - \beta_\phi \mu \exp(\beta_\phi \mu)}{\exp(\beta_\phi \mu) - \kappa_\phi}\right)} (\kappa_\phi - \exp(\beta_\phi \mu))}{(\kappa_\phi - 1)}. \tag{1.28}$$

After acquiring κ_ϕ, μ and $\widetilde{n}_\phi^0(t_{[\beta_\phi,\infty]})$ experimentally, this expression will depend on a single unknown β_ϕ. One can show that Eq. 1.28 is solved by a unique β_ϕ. This follows from the fact that the right hand side of Eq. 1.28 is a monotonically decreasing function in $\beta_\phi \in [0, \ln(\kappa_\phi)/\mu]$ with corresponding values lying in the interval $[\widetilde{n}^{\phi} \exp(-\frac{\kappa_\phi \mu t_{[\beta_\phi,\infty]}}{\kappa_\phi - 1}), 0]$ while the left hand side is positive by definition. Substituting the solution for β_ϕ into Eq. 1.14 yields the remaining parameter vector α_ϕ.
Taken together this proves that in theory samples of the three cell cohorts taken at three support points, a first at $t = 0$, a second at $0 < t < \min(\boldsymbol{\beta})$ and a third at $t > \max(\boldsymbol{\beta})$ are sufficient to determine all the parameters of the model.

1.4.4 Validation with DNA-BrdU pulse-chase labeling experiments

The *Gedankenexperiment* analyzed so far, although conceptually simple, poses a series of experimental challenges, that make a one-to-one realization difficult. The technical difficulties lie mostly in initially separating the cells according to their phase and in following these cells as they enter the subsequent phases. A widely used technique, namely DNA-nucleoside pulse-chase labeling experiments, generates nevertheless to a certain extent comparable data. The latter achieves, as discussed in Section 1.2, the initial phase-specific partitioning by exposing, during a short time window, proliferating cells with a thymidine analog (e.g., BrdU) that gets selectively incorporated into the DNA of cells that are actively replicating their genome. Measuring subsequently by FACS simultaneously the DNA content and the amount of incorporated thymidine analog permits to discern the three phases G_1, S and G_2M immediately after the pulse. In addition, due to the permanent staining property of the thymidine analogs, it is possible to follow, up to a certain degree, the labeled and unlabeled cell cohorts over time (for more details see Section 1.2).
In theory this method would largely correspond to the hypothetical experiment that we analyzed so far. In practice however, the overlap of the subpopulations in the FACS scatter plots prevents the exact determination of the frequencies of cells described by Eq. 1.25. For example labeled cells that have completed the S phase but remain in G_2M phase are indistinguishable from those that did not complete the initial S phase yet. As has been reported previously, only four different sub-populations can be identified with reasonable accuracy [14]. These are:

- f^{lu} : labeled undivided cells which at time of labeling (t_0) were in S phase ($n_0^S + n_1^S$),
- $f^u_{G_2M}$: unlabeled cells that were in G_2M phase at t_0 ($n_0^{G_2M}$),
- f^{ld} : labeled divided cells which were initially in S phase ($\sum_{p=2}^{4} n_p^S$) ,
- $f^u_{G_1}$: unlabeled cells and progeny of cells that were in G_1 at t_0 accompanied by the progeny of $f^u_{G_2M}$ and f^{lu} ($1 - f^{lu} - f^u_{G_2M} - f^{ld}$),

where the corresponding populations in our *Gedankenexperiment* are indicated in parenthesis. This shows that computing Eq. 1.26 up to p equal 4 is sufficient to describe a complete *in silico* BrdU

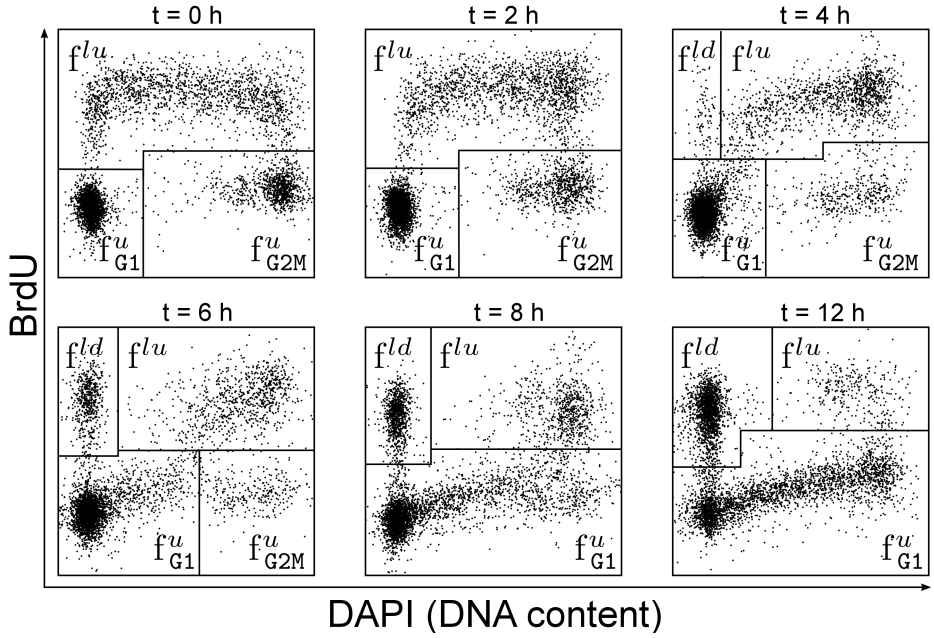

Figure 1.3: DAPI-BrdU pulse-chase labeling FACS data. Samples taken at several time points after pulse labeling proliferating U87 human glioblastoma cells with BrdU (for details see Materials and Methods, Experimental Methods). The four gated populations are f^{lu}, $f^{u}_{G_2M}$, $f^{u}_{G_1}$ and f^{ld} defined in the main text. Note that for this data none of the cells initially in S apparently divided twice. The experiment was designed in collaboration with and performed by Dr. Jaehnert at Ludwig-Maximilians University in Munich.

pulse labeling experiment. The reason is that using current protocols BrdU labeling becomes indistinguishable from background as soon as the latter divide a second time. In other words, cells that leave population f^{ld} by dividing a second time join population $f^{u}_{G_1}$ (see Fig. 1.3). It is worth noticing that in the case of the experimental data analyzed in the next section there is no evidence that labeled cells divided twice during the experiment.
The population $f^{u}_{G_2M}$ is the only sub-population that matches directly the type of data considered before and its temporal evolution follows as such Eq. 1.25. The remaining three populations in contrast represent mixtures of cell cohorts whose kinetics could be described individually by Eqs 1.25-1.26.

Learning from real DNA-nucleoside analog pulse-chase labeling experiments

By analyzing two data sets from samples of BrdU pulse-chase labeling experiments, we tested the model and the effect of population intermixing on the identification of the model parameter values. The two cell lines considered were *in vitro* cultured U87 human glioblastoma cancer cells (for details see Materials and Methods) and *in vitro* grown V79 Chinese hamster cells (courtesy G. Wilson). We will refer to these data as the U87 and the V79 data sets. The two latter consist of samples taken at several time points after a single BrdU pulse, with sample sizes ranging from 5000 to 50000 cells each. Data points represent simultaneous measurements of BrdU as well as DAPI or PI (DNA content) in single cells by fluorescent activated cell sorting. The U87 data set was generated in collaboration with Christian Schichor and Irene Jaehnert at the Department of Neurosurgery, Ludwig-Maximilians-University Munich, Klinikum Grosshadern.
As a preliminary test, we minimize the residual sum of squares (RSS), i.e., least-squares fitting, of

adequate mixtures of Eq. 1.25 and Eq. 1.26 to extracted frequencies at different time points after the pulse. We find that, for properly chosen parameter values, both data sets are reasonably well approximated by the model predictions (see Fig. 1.4 A).
While this indicates that the model captures some of the relevant temporal characteristics of cell cycle progression, a more careful analysis reveals that an infinite number of different parameter combinations does fit the measured frequencies with the same minimal RSS (not shown). This implies that there exists, given the available data, no single best-fit parameter combination, but a whole region in parameter space which can explain the data equally well.
When we then interrogate the same data by approximate maximum likelihood (ML) estimation, using a simple *ad hoc* likelihood function Eq. 1.33 (see Material and Methods, Bayesian Inference), we find again that relative large regions in parameter space map to the same ML (see Fig. 1.5). It turns out that these regions are entirely superimposed onto the lines defined by Eq. 1.14 and shown in Fig. 1.5 (dashed lines). The latter delineate what could have potentially been learned in our *Gedankenexperiment* with only two support points, namely one at $t = 0$ and a second at $t < \min(\boldsymbol{\beta})$. In both experiments, ML parameters associated with the G_1 phase are spread out almost everywhere along these lines. Parameters related to the S phase are more concentrated but still in the case of the V79 data a substantial region of ML estimates is observed. Finally the region for the G_2M phase parameters approaches that of a point estimate for both data sets.
The spread of the ML estimates suggests that even in the ideal case of noise-free data, the specific choice of the support points in these experiments does not allow to determine uniquely the variability neither for all the phases nor for the total cell cycle length distribution. In contrast the average completion time for each phase, corresponding approximately to the width and not to the length of the regions, can still be estimated with relative high precision.
To further quantify the uncertainty of these estimates, Bayesian 99%-credibility regions (CR) are computed with the Markov chain Monte Carlo method (see Appendix, Algorithm) using the same likelihood function as before (Fig 1.6). CRs follow mainly the same trends as the regions observed in the ML estimates, cover however as expected a larger volume. An exception is the 'blown up' CR of the S phase parameter for the U87 cell line, for which the ML estimates wrongly insinuate a well defined point estimate.
In Table 1.1 we summarize the obtained Bayesian summary statistics. One can see, in line with our balanced growth analysis, that the intervals for the average duration of each phase are narrow compared to those for the individual parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. In both cases the data allows for a deterministic S phase ($\alpha = 0$), while for the U87 data set variability in G_2M is a necessary characteristic to reproduce accurately the data. Notably, when contrasting the two cell lines, are the short G_1-phase of Chinese hamster cells and the approximately two times more extended G_2M-phase of the glioblastoma cell line. It is out of the scope of this thesis to interpret or relate these differences to cell line specific conditions. More important in this context is the fact that the information contained in the analyzed data is too sparse to narrow down all the parameter values even under noise-free conditions. We will address this issue in the next chapter, where improving the performance of pulse-chase labeling experiments will be our main target.

1.5 Materials and Methods

1.5.1 Stability analysis

From a parameter estimation perspective the results derived in this chapter are only useful, if balanced growth corresponds to a situation that is experimentally relevant *in vitro* or *in vivo*. Empirically, exponential growth of proliferating cell populations is routinely observed in cell cultures under ideal conditions, given virtually unlimited space and nutrients. Even if these conditions are necessarily temporary, there usually exist a time window during which balanced growth may represent a good

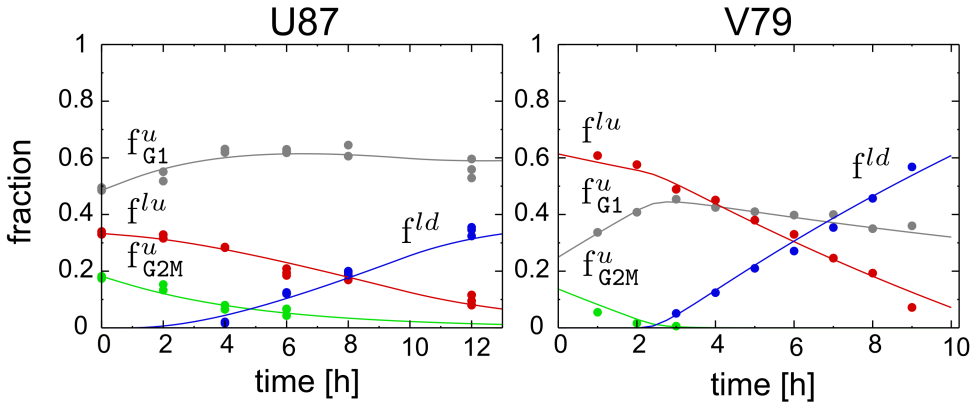

Figure 1.4: Least-squares model fitting. Best fit of the model predictions (lines) to experimentally determined cell fractions after BrdU pulse labeling (dots). U87: In vitro cultured U87 human glioblastoma cancer cell line (see Fig. 1.3). V79: In vitro cultured V79 Chinese hamster cells (courtesy G. Wilson).

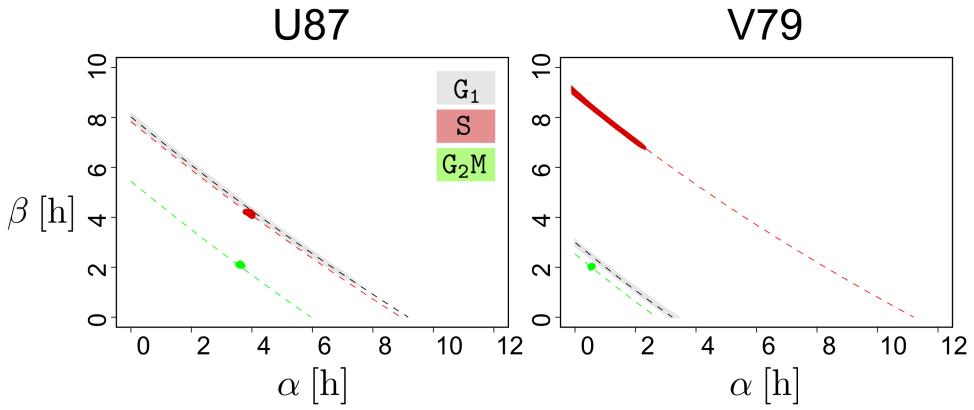

Figure 1.5: Maximum likelihood estimation. Bi-variate maximum likelihood regions for the parameters α and β associated to each phase (gray: G_1, red: S, green: G_2M). The dashed lines delineate the information that could have been gained in our *Gedankenexperiment* from two support points, one at $t = 0$ and a second at $t < \min(\boldsymbol{\beta})$.

	G_1 [h]			S [h]			G_2M [h]			$\bar{T}$ [h]
	α	β	$\bar{\tau}$	α	β	$\bar{\tau}$	α	β	$\bar{\tau}$	
U87	4.8	4.1	8.9	3.3	4.8	8.2	3.6	2.0	5.6	22.9
	(0.0:11.2)	(0.0:9.7)	(6.7:11.7)	(0.0:5.9)	(2.0:8.8)	(7.1:9.5)	(2.7:4.6)	(1.3:2.7)	(5.1:6.2)	(19.4:26.4)
V79	1.6	1.5	3.1	1.3	7.6	9.0	0.5	2.0	2.5	14.7
	(0.0:3.6)	(0.0:3.4)	(2.5:3.8)	(0.0:3.6)	(5.5:9.6)	(8.4:9.7)	(0.2:0.7)	(1.7:2.3)	(2.3:2.7)	(13.7:15.9)

Table 1.1: Bayesian summary statistics (mean, 99%-credibility intervals) for individual cell cycle parameters, average durations $\bar{\tau} = \alpha + \beta$ and the total cell cycle length $\bar{T}$. The intervals for the average durations are narrow compared to those for the individual parameters α and β.

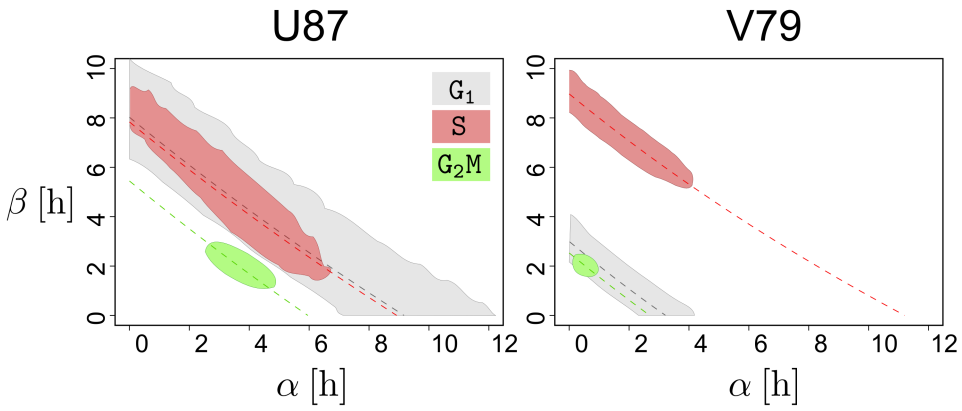

Figure 1.6: Bayesian Inference. Bi-variate 99%-credibility regions for the parameters α and β associated to each phase (gray: G_1, red: S, green: G_2M). CRs follow mainly the same trends as the regions observed in the ML estimates (see Fig. 1.5), cover however a larger volume. An exception is the 'blown up' CR of the S phase parameter for the U87 cell line, for which the ML estimates wrongly insinuate a well defined point estimate.

approximation. This, for example, is reflected in the logistic growth model, whose kinetics in a first phase are indistinguishable from exponential growth. Here we will show on a theoretical basis that a population that follows the stochastic model specified before, will eventually enter this so-called exponential growth phase.

The requirement for such an asymptotic behavior is, recalling Eq. 1.13, that, for phase $i \in \{G_1, S, G_2M\}$, the complex valued function

$$Q(\mu) = 2 - \prod_i e^{\beta_i \mu}(1 + \alpha_i \mu) \tag{1.29}$$

has for given positive α_i and β_i a unique positive real root which represents the upper bound of the real part of any of its other potentially infinite number of roots. The complex number μ that solves Eq 1.29 corresponds, according to our model, to the stationary phase growth rate of the proliferating cell population. In case that μ is real, the population is growing exponentially, while if μ is purely imaginary, growth is oscillatory. In general, roots have both non-vanishing real and imaginary parts, which leads to oscillations with growing or decaying amplitude.

If for real x and y we write $\mu = x + iy$, the real and imaginary part of $Q(\mu)$ are computed as

$$\begin{aligned} \mathrm{Re}(Q) &= 2 - e^{Bx}(\psi_1 \cos(By) - \psi_2 \sin(By)), \\ \mathrm{Im}(Q) &= -e^{Bx}(\psi_2 \cos(By) + \psi_1 \sin(By)), \end{aligned} \tag{1.30}$$

where

$$\begin{aligned} B &= \sum \beta_i, \\ \psi_1 &= \prod_i (1 + \alpha_i x) - y^2 \left(\sum_i \alpha_i^{-1} + 3x \right) \prod_i \alpha_i, \\ \psi_2 &= y \left(\sum_i \alpha_i + (2 \sum_i \alpha_i^{-1} + 3x^2 - y^2) \prod_i \alpha_i \right). \end{aligned} \tag{1.31}$$

For μ to be a root of Q both real and imaginary part have to vanish. We restrict our analysis to the positive complex half plane, i.e. $x \geq 0$, because we are interested in growing and not contracting cell populations. Due to the symmetries in the trigonometric functions $\sin(-y) = -\sin(y)$ and $cos(-y) =$

$cos(y)$ and $\psi_1(-y) = \psi_1(y)$ and $\psi_2(-y) = -\psi_2(y)$ one can see that if $\mu = x + iy$ is a root, its complement $\mu^* = x - iy$ is also a root. We can thus reduce the analysis even further to values with positive imaginary parts. If for fixed x we plot Q in the complex plane as a parametric function of $y \in [0, \infty]$ we get a spiral with the distance from a real center point $c = 2 + i\,0$ given by

$$\begin{aligned} r =& \sqrt{(\mathrm{Re}(Q) - 2)^2 + \mathrm{Im}(Q)^2}, \\ =& e^{\omega x} \sqrt{\prod_i ((1 + \alpha_i x)^2 + \alpha_i^2 y^2)}. \end{aligned} \tag{1.32}$$

Crucially, as r is a monotone increasing function in y, the spiral never crosses itself. For $y = 0$ the imaginary part of Q vanishes as expected because $\lim_{y\to 0} \sin(\omega y) = 0$ and $\lim_{y\to 0} \psi_2 = 0$. For this special case, $\mathrm{Re}(Q) = 2 - \prod_i e^{\beta_i x}(1 + \alpha_i x)$ is obviously monotone decreasing with increasing x and restricted to the interval $[1, -\infty]$. This means that the spiral can only 'start' in the interval between one and minus infinity. Taken together this implies that if for $y = 0$ and fixed x, the real part of Q is positive, there exist a single 'opportunity' to cross the origin, while if the real part of Q is negative there exists none. At the border where the real part is zero (Fig. 1.7 C), the corresponding value of x is the only positive real root. Due to the monotonicity of $\mathrm{Re}(Q)$ any value of x greater than the positive real root will result for $y = 0$ in $\mathrm{Re}(Q) < 0$ which does not admit for any solution. The different possible scenarios are exemplified in Fig. 1.7.

1.5.2 Bayesian inference

When estimating, by FACS analysis, frequencies of cells in different phases of the cell cycle, measurement noise becomes unavoidable. Potential sources of noise include variability in experimental conditions, gating errors, stochasticity in cell division, FACS measurement errors, and many more. Here we describe an attempt to account, in a simple way, for the observed experimental noise by taking a Bayesian approach. This provides us not only with a maximum likelihood estimate region of the model parameter, but in addition will give us an idea about the uncertainty that we have about the parameter values.

Even though considering all potential sources of noise would be most consistent, the resulting probability model would become far more complex than our initial cell cycle model. To avoid this overload we assume that a relatively simple *ad hoc* multivariate probability density function approximates reasonably well the average and the noise in the observed frequencies at a single time point. This probability density function, which corresponds to the likelihood $\mathcal{P}_i$ of a single measurement event $\widetilde{\boldsymbol{n}}_{\boldsymbol{i}}$, is defined by

$$\mathcal{P}_i(\widetilde{\boldsymbol{n}}_i|\boldsymbol{\alpha}, \boldsymbol{\beta}, N; t_i) = \Gamma(N) \prod_{j=1}^{k} \frac{\widetilde{n}_{i,j}^{N n_{i,j}-1}}{\Gamma(N n_{i,j})}, \tag{1.33}$$

where Γ is the Euler gamma function. The right-hand side of Eq 1.33 corresponds to a continuous approximation of a scaled multinomial distribution with support $x_j \in [0, 1]$ and $\sum x_j = 1$ [90]. The parameter N, which determines the spread of the distribution, can be interpreted as an effective population size. Taking e.g., a sample of size N from a population of cells containing k sub-populations with proportions given by n_i yields frequencies with a probability density approximately distributed accordingly. If N is small the density distribution is broad, while if N becomes large the density distribution becomes narrow.

Following in general terms the notation in the main text, the $\widetilde{n}_{i,j}$ denote the k measured population frequencies from experiment i and the $n_{i,j}$ stand for the corresponding frequencies predicted by the cell cycle model. The latter obviously depend on the parameter vector $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and the time t_i.

Having defined the likelihood $\mathcal{P}_i$ for an outcome of a single pulse labeling experiment, the likelihood for

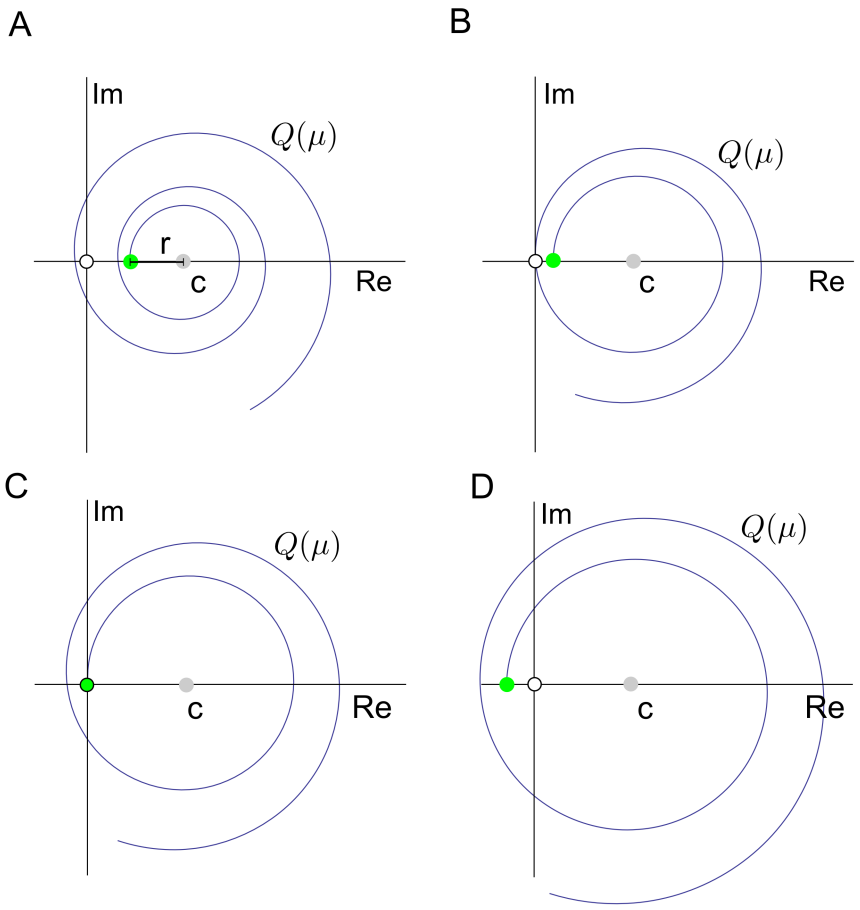

Figure 1.7: Here we plot $Q(x+iy)$ as a function of $y \geq 0$ for fixed values of $x \geq 0$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For $y = 0$ (green circle) the real part of Q takes, depending on $x \in [0, \infty]$, a value in the interval $[1, -\infty]$. The values for x are increasing from A-D, while $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ remain unchanged. For relatively low values of x (A-B) the real part $\mathrm{Re}(Q)$ is positive for $y = 0$. After one or several turns, i.e by increasing y the spiral can potentially cross the origin only once (empty circle). In A the spiral misses the origin, while in B the spiral crosses the origin after one turn. Crossing of the origin means that the corresponding complex number $\mu = x + iy$ is a root of Q. In C the spiral starts at the origin. This represents the only real positive root of Q. For initially negative values of $\mathrm{Re}(Q)$ (D) the spiral can never cross the origin because the distance to the center point (gray circle) is already in the beginning for $y = 0$ larger than the distance between the latter and the origin. By increasing y this distance will even grow further according to Eq. 1.32 .

the outcomes of a set of m experiments is the product $\mathcal{P} = \prod_{i=1}^{m} \mathcal{P}_i$ under the reasonable assumption that noise in a specific experiment is independent of all the other experiments. By numerically inverting $\mathcal{P}$, using Bayes theorem, one can obtain the posterior and subsequently the uncertainty over the model parameter given the data, the model and prior knowledge.
To estimate the maximum likelihood regions, the posteriors and the uncertainties in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for the U87 and V79 data sets, we implement the adaptive Markov-Chain-Monte-Carlo algorithm proposed in [91] (see also Appendix, Algorithm). The estimates for the maximum likelihood regions are obtained by fixing N to an extreme value (e.g., $1e^5$). For Bayesian inference, N is considered as an additional parameter. For simplicity, improper priors uniformly distributed over the positive real number are assumed for all parameter. The first 10^6 steps of the initially 10^7 step-long chains are discarded, and of the remaining chains every 1000'th step is included in the subsequent analysis. The credibility regions are computed from the resulting MCMC chains using the 'HPDregionplot' routine in the R package 'emdbook', and convergence of the chains is confirmed using the Gelman convergence test.

1.5.3 Some useful identities

In this section, several identities will be derived, which turn out to be useful in generalizing the model solutions to arbitrary waiting time distributions with known expectation and known Laplace transforms. These identities are as follows:

$$\int_{-\infty}^{t} f(t-x)dx = 1, \tag{1.34}$$

$$\int_{-\infty}^{t} R(t-x)dx = E\{f(x)\}, \tag{1.35}$$

$$\int_{-\infty}^{t} e^{\mu x} f(t-x)dx = e^{\mu t} \times \mathcal{L}_\mu\{f(x)\}, \tag{1.36}$$

$$\int_{-\infty}^{t} e^{\mu x} R(t-x)dx = e^{\mu t} \times \frac{1-\mathcal{L}_\mu\{f(x)\}}{\mu}, \tag{1.37}$$

$$\int_{-\infty}^{0} f(t-x)dx = R(x), \tag{1.38}$$

$$\int_{-\infty}^{0} R(0-x)dx = E\{f(x)\}, \tag{1.39}$$

$$\int_{-\infty}^{0} e^{\mu x} f(t-x)dx = \mathcal{L}_\mu\{f(x+t)\}, \tag{1.40}$$

$$\int_{-\infty}^{0} e^{\mu x} R(0-x)dx = \frac{1-\mathcal{L}_\mu\{f(x)\}}{\mu}, \tag{1.41}$$

where $\mathcal{L}_\mu\{f(x)\} = \int_0^\infty e^{-\mu x} f(x)dx$, is the Laplace transform, $R(t) = 1 - \int_0^t f(x)dx$, is the survival or reliability function, and $E\{\cdot\} = \int_0^\infty x f(x)dx$, is the expectation. As most identities follow directly by substitution of $(t-x)$ with y, we will indicate, for brevity, an equality which uniquely involves this

specific operation, by the symbol $\stackrel{s}{=}$.
For Eq. 1.34 we have

$$\int_{-\infty}^{t} f(t-x)dx \stackrel{s}{=} -\int_{\infty}^{0} f(y)dy = \int_{0}^{\infty} f(y)dy = 1,$$

where the last step is a consequence of the fact that $f(y)$ is a normalized probability density function for a non-negative random variable.
For the left hand side of Eq. 1.35 we can write

$$\int_{-\infty}^{t} R(t-x)dx \stackrel{s}{=} \int_{0}^{\infty} R(y)dy.$$

To show that this last expression equals the expectation of $f(t)$, we let

$$\begin{aligned} u(y) &= y, \qquad & v(y) &= -R(y), \\ u'(y) &= 1, \qquad & v'(y) &= f(y), \end{aligned}$$

and using integration by parts (i.e., $\int_a^b uv' = uv|_a^b - \int_a^b u'v$) we get

$$\int_0^b yf(y)dy = -yR(y)\big|_0^b + \int_0^b R(y)dy.$$

From this, recalling the definition of the expected value, and recognizing that the first term on the right hand size approaches zero as $b \to \infty$, we obtain the classical result

$$E\{f(y)\} = \int_0^\infty R(y)dy,$$

which confirms the identity in Eq. 1.35.
The identity in Eq. 1.36 is derived straightforwardly as follows

$$\int_{-\infty}^{t} e^{\mu x} f(t-x)dx \stackrel{s}{=} \int_{0}^{\infty} e^{\mu(t-y)} f(y)dy = e^{\mu t} \times \mathcal{L}_\mu\{f(y)\}.$$

Similarly, we can rewrite the left hand side of Eq. 1.37 as

$$\begin{aligned} \int_{-\infty}^{t} e^{\mu x} R(t-x)dx \stackrel{s}{=} \int_{0}^{\infty} e^{\mu(t-y)} R(y)dy &= e^{\mu t} \times \mathcal{L}_\mu\{R(y)\}, \\ &= e^{\mu t} \times \mathcal{L}_\mu\{1 - \int_0^y f(y)dy\}, \\ &= e^{\mu t} \times \mathcal{L}_\mu\{\frac{1}{\mu} - \frac{\mathcal{L}_\mu\{f(y)\}}{\mu}\}, \\ &= e^{\mu t} \times \frac{1 - \mathcal{L}_\mu\{f(y)\}}{\mu}, \end{aligned}$$

where in the third row, we exploited the well known property of the Laplace transform which says that

the Laplace transform of a given cumulative function is the Laplace transform of its density function, divided by the transformed variable, which in this case is μ.
For Eq. 1.38 we have

$$\int_{-\infty}^{0} f(t-x)dx = 1 - \int_{0}^{\infty} f(t-x)dx \stackrel{s}{=} 1 + \int_{t}^{-\infty} f(y)dy$$
$$= 1 - \int_{-\infty}^{t} f(y)dy = 1 - \int_{0}^{t} f(y)dy = R(t),$$

in which the first step follows from $\int_{-\infty}^{\infty} f(t-x)dx = 1$, and the second last equality uses the fact that $f(y)$ is zero for $y < 0$.
Eq. 1.39 is found by substituting t by zero in Eq. 1.35.
To show Eq. 1.40, we write

$$\int_{-\infty}^{0} e^{\mu x} f(t-x)dx = \int_{0}^{\infty} e^{-\mu z} f(t+z)dz = \mathcal{L}_{\mu}\{f(z+t)\},$$

where we substituted x by $-z$ to derive the first equality. The last equality ensues directly from the definition of the Laplace transform.
Finally, Eq. 1.41 is obtained by substituting t by zero in Eq. 1.37.

1.5.4 Experimental methods

Cell culture experiments and BrdU labeling protocols for the U87 data set were designed in collaboration with Christian Schichor and Irene Jaehnert at Department of Neurosurgery, Ludwig-Maximilians-University Munich, Klinikum Grosshadern. For the sake of completeness, experimental methods are provided below, according to Irene Jaehnert, who performed the actual experiments.

Cell culture

Human astrocytoma cells U87 MG (ATcell cycle-LGC) were routinely cultured with Dulbecco's modified Eagles medium (DMEM, Biochrom AG) supplemented with non essential amino acids (NEAA, Invitrogen GmbH), heat-inactivated fetal bovine serum (FBS, 10%, Biochrom AG) and additives (penicillin-streptomycin-glutamine, Invitrogen GmbH) in plastic flasks (TPP AG) at 37 °C in 5% CO2-humified incubators and were passaged twice a week using Dulbecco's PBS (DPBS, Apotheke Innenstadt Uni München) and Trypsin/EDTA (Biochrom AG) before reaching confluence.

Treatment with BrdU

For cell cycle analysis cells (2.0 x $10^4/cm^2$) were seeded in 75 cm2 culture flasks and incubated for 24 h followed by the BrdU pulse. For this purpose, medium was replaced by medium supplemented with BrdU (10 µM, Bromodeoxyuridine, Becton Dickinson GmbH), cells were incubated for 30 min at 37 °C followed by washing away of BrdU for two times with fresh medium. Cells were then again incubated at 37 °C for a designated period of time (0 h, 2 h, 4 h, 6 h, 8 h, 12 h) to measure proliferation over 12 h.

Preparation of samples

Collecting of cells was performed by trypsinization using DPBS, Trypsin/EDTA and medium followed by washing of cells in DPBS. To exclude dead cells from the analysis staining of dead cells was performed. For this purpose cells were incubated for 30 min with fluorescent dye (LIVE/DEAD Fixable Green Dead Cell Stain Kit, Invitrogen) according to the manufacturers instructions followed by washing with DPBS. Consequent steps of sample preparation were processed using the APC BrdU Flow Kit (Becton Dickinson GmbH). Cells were washed once with Perm/Wash Buffer and fixed for 30 min on ice with Cytofix/Cytoperm Buffer. After washing with BD Perm/Wash Buffer cells were resuspended in Cytoperm Plus Buffer and incubated on ice for 10 min followed by washing with Perm/Wash Buffer and incubation in Cytofix/Cytoperm Buffer for 5 min on ice. Cells were then washed with Perm/Wash Buffer and incubated with 2 M HCl-Triton (1%) for 30 min at room temperature followed by washing twice with Perm/Wash Buffer. For detection of incorporated BrdU cells were incubated with diluted (1:50) fluorochrome-conjugated anti-BrdU antibody for 20 min at room temperature. Cells were then washed with BD Perm/Wash Buffer and further incubated with DAPI (0.5 µg/ml in staining buffer: 100 mM Tris, pH 7.4, 150 mM NaCl, 1 mM $CaCl_2$, 0.5 mM $MgCl_2$, 0.1% Nonidet P-40) for 30 min at room temperature. All samples have subsequently been stored on ice until acquisition.

Acquisition and analysis

Acquisition of data was performed by measuring fluorescence intensity using a BD LSR II Cytometer at the excitation wavelength of 660 nm for APC and 450 nm for DAPI and the software BD FACSDiva.

1.6 Discussion

In this chapter we analyzed a simple stochastic model that aims at approximating the time it takes for a cell to accomplish the sequential phases of the cell cycle, by defining the completion time in each phase as a shifted exponential density distribution. At first sight this might seem a gross oversimplification of all the processes involved. However, when compared with experimental data, this simplistic model performs surprisingly well.

While the observation that the model reproduces well the experimental time series has to be interpreted with care, we think its performance can be accounted to the fact that the probability rule captures simultaneously two important regimes of complex biochemical processes, that qualitatively differ in their completion time distribution. As was shown recently by Bel *et* al. [92] the completion time for a large class of complex theoretical biochemical systems, including models for DNA synthesis and repair, protein translation and molecular transport, simplify either to deterministic or to exponentially distributed completion times, with a very narrow transition between the two regimes. These are precisely the 'ingredients' of the shifted exponential distribution. Under this light our model could be naively interpreted as a sensor that measures approximately the 'relative contribution of delay and decay processes' in each of the cell cycle phases. However, while in series connected delays form again a delay, this is not true for decays. These form a process with hypoexponential distributed completion times with a shape similar to the frequency distribution of the S phase completion time reported in [72]. Thus a more flexible model for the completion time of each phase could be a hypoexponential distribution, a distribution of the family that we are currently using to model the total cell cycle length (i.e., Eq. 1.5). Or alternatively, if processes are not in series connected but rather concurrent, the time for all the processes to complete is dominated by the largest delay or the smallest decay parameter. This might explain the close to negative-exponentially distributed completion time for G_1 which was measured by the technique of cell cycle specific reporter genes [72].

An important simplification of our model consist in the assumption that cell loss is small compared to population wide division rates, such that we can neglect apoptotic and necrotic cells when we fit the model to experimental data. The main reasons to adopt this approach is simplicity and the fact

that the available data sets does hardly permit the determination of the possibly large number of additional parameters. While for the U87 LIFE/DEAD discrimination was performed, the markers used for gating are specific for late stages of apoptosis or necrosis, typically after membrane integrity is lost, and therefore do not necessarily reflect the true fraction of dying cells. The fraction of dead cells identified and excluded by this method was typically low. In case that experimental conditions would however suggest a substantial death rate, the model is flexible enough to be adapted without major technical difficulties. Given that the apoptotic state (e.g., defined by Annexin V staining) would be measured simultaneously with BrdU and DNA content, this could even open up the possibility to assess the stochastic timing of apoptosis *in vivo*.

Another fundamental abstraction of our model is found in the presumption that the completion times for the cell cycle phases of a given cell are uncorrelated, which also implies uncorrelated division times of parental cells and siblings. Even though positive correlation in division times between parental and daughter cells [42] and between siblings [93] has been observed recently *in vitro* by direct long-term microscopy of activated proliferating B cells, Schultze *et* al. showed many years ago for *in vivo* murine crypt epithelial cells the lack of correlation of completion times of a cell through successive phases [**?**]. It remains to be shown experimentally how much of the correlation or lack of correlation is due to cell type or environment. In any case it would be interesting to adapt the present model such that correlation of phase completion times could be accounted for in the estimation procedure.

After investing a considerable amount of effort into developing and solving a stochastic cell cycle model, the awaited output from experimental data remained relatively disappointing. Even thought the model reproduces well the empirical kinetics, many of the parameters from the cells under study are only partially identified. This could indicate that we have overparameterized the system, however two parameters do not seem disproportionate to describe the completion time distributions of a cell cycle phase. Alternatively, the experimental setup might not have been ideal to validate the model. This is the hypothesis, that we will pursue in the next chapter, where the model developed herein will be used to derive experimental designs which are optimal in a sense that they help to identify the model parameters most efficiently.

2 Improving the Design of DNA-Nucleoside Pulse-Chase Labeling Experiments

2.1 Motivation and Background

The two ultimate goals of many model-based inference strategies are : (i) model validation and (ii) parameter estimation. While the outcome of (i) mostly depends on how well the often highly simplified model captures measurable traits of the underlying more complex 'real' process, the success of (ii) requires that the information contained in the collected data is sufficient to identify the parameter values. The less data is analyzed, the more likely is (i) but the more difficult becomes (ii). The most common reason why (ii) is not satisfactory is measurement noise or insufficient sampling, however other possibilities exist. For instance when the kinetics predicted by a model depend on quantities that are not directly visible, e.g., hidden Markov models, uniquely identifying certain parameters may turn out to be impossible [94]. Similarly, solutions of a model may not be invertible and parameter values may remain unidentifiable due to the non-injectivity of the model's parametrization map [95].
We concluded the first chapter of this thesis with the observation that even though our model could reasonably well reproduce empirical cell cycle kinetics, some of the parameters in the model remained undetermined. This indicated successful (i)[1] but partial failure in (ii). In this chapter we will investigate the deeper reason for this failure, and analyze *in silico* ways how to avoid this situation in future studies. One approach will exploit the theory behind the design of experiments (DoE) in order to optimize sampling schemes in pulse labeling experiments, while a second approach will be based on a dual pulse labeling protocol.
Before discussing our main findings however, basic insights from the field of experimental design are briefly summarized and former dual pulse labeling studies are reviewed.

2.1.1 Optimal design of experiments

Design of experiments (DoE) is an essential part of any data driven research. From a simplistic point of view, it consists in a first step, in choosing the type of experiment that is most appropriate to test a given hypothesis or to estimate a certain quantity of interest. Then, in a second step, ignoring other technical details, the number of samples, the position of the support points and the distribution of the samples over the support points have to be specified. Both steps are usually subject to monetary, temporal and technical constraints. Therefore, provided that the first step has been concluded, an ideal second step would consist in choosing a sampling scheme that maximizes, under the above mentioned constraints, the information in the data concerning the research question at hand. In general, such an optimal experimental protocol is not known and common sense, experience or intuition are applied instead. If however a mathematical model is available that is able to describe reasonably well the expected data, this model can be exploited to find an optimal sampling scheme. In the following, we will review the basic theory that has been developed over the last two hundred years to solve this optimization problem [96, 97].
As a simple, yet informative example, we first consider the design problem for linear models of the

[1] Notice that, according to Sir Karl Raimund Popper and to common sense, a hypothesis, i.e., a model, can only be disproven but never proven. Success in this case means that a given hypothesis has not been disproven by the data.

form

$$\mathbf{y} = \boldsymbol{X}(\boldsymbol{\xi})\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is a $n \times 1$ matrix of n uni-variate observations, $\boldsymbol{\theta}$ is a $p \times 1$ matrix of p parameters, $\mathbf{X}(\xi)$ is a $n \times p$ matrix whose entries depend on the design $\boldsymbol{\xi}$, and $\boldsymbol{\epsilon}$ is an $n \times 1$ matrix of n independent errors which are normally distributed; $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ for $i = 1, 2, \ldots, n$. Usually it is assumed that the number of observations equals the number of parameters, in other words is is required that n equals p.

As a special case, consider a quadratic response-surface model with one explanatory variable (e.g., temperature, time), three parameters and three observations :

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & \xi_1 & \xi_1^2 \\ 1 & \xi_2 & \xi_2^2 \\ 1 & \xi_3 & \xi_3^2 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.$$

While the design $\boldsymbol{\xi}$ specifies for each observation the controllable experimental conditions (e.g., the temperature of the system, the time point of measurement), the matrix $\mathbf{X}(\boldsymbol{\xi})$ describes the functional relationship between these conditions and the expected data.

A classical result for linear models, as defined above, states that the least-squares estimator $\widehat{\boldsymbol{\theta}}$ based on a realization of $\mathbf{y}$ is distributed according to $\mathcal{N}(\boldsymbol{\theta}, \boldsymbol{M}^{-1})$, where $\boldsymbol{M} = \mathbf{X}'\mathbf{X}$ is the Fisher information matrix [98]. From this follows that the volume of the joint confidence ellipsoid for $\widehat{\boldsymbol{\theta}}$ is proportional to $(\det(\boldsymbol{M}))^{-1/p}$. Therefore, if the purpose of an experiment is to estimate the parameters as accurately as possible, a reasonable design criterion could aim at maximizing $\det(\boldsymbol{M})$ [99–101]. This criterion is known as the determinant criterion and a design which maximizes $\det(\boldsymbol{M})$ is called D-optimal. Other related, however less commonly applied criteria minimize for example the trace of $\boldsymbol{M}^{-1}$ (A-optimal) or the maximal eigenvalue of $\boldsymbol{M}^{-1}$ (E-optimal) [102].

Importantly $\boldsymbol{M}$ and therefore also D-optimal designs for linear models do not depend on the parameter $\boldsymbol{\theta}$. This is not true for non-linear models, which are typically defined as follows

$$\mathbf{y} = \eta(\boldsymbol{\xi}, \boldsymbol{\theta}) + \boldsymbol{\epsilon},$$

where the expected response $\eta(\boldsymbol{\xi}, \boldsymbol{\theta})$ contains non-linear functions of the parameters such that it cannot be expressed in matrix form as before. Now the Fisher information matrix is computed as

$$\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \frac{\partial \eta(\xi_i, \boldsymbol{\theta})}{\partial \theta} \frac{\partial \eta(\xi_i, \boldsymbol{\theta})}{\partial \theta^T}, \tag{2.1}$$

and the determinant criterion could in principle be applied as before. However, because $\boldsymbol{M}$ depends on $\boldsymbol{\theta}$, which is typically to be estimated, choosing a D-optimal design is not as straightforward.

As simple approximate solution to this problem is to use a 'best guess' estimate, say $\hat{\boldsymbol{\theta}}_0$, possibly derived from a previous experiment, and optimize as before. This procedure yields locally D-optimal designs, where the term 'locally' emphasizes the fact that the optimal solution is based on an initial parameter choice [103]. Alternatively, if a prior distribution $p(\boldsymbol{\theta})$ over the parameter $\boldsymbol{\theta}$ is available, Bayesian D-optimal designs can be derived which maximize

$$S = \int \log\{\det(\mathbf{M}(\boldsymbol{\xi}, \boldsymbol{\theta}))\} p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{2.2}$$

which corresponds to maximizing the expected gain in Shannon information [104].

If experiments are expensive, one might wish to use the information obtained in a first part of the experiment, in order to optimize the design in a second part. This approach is called batch-sequential

experimental design, where the batch size gives the number of new support points that are sampled before a new optimization round is performed [105]. The extreme case in which the number of support points is increased one at a time is known as fully sequential D-optimal experimental design [106]. Here new design points are sequentially added such that

$$\xi_{k+1} = \arg\max_{\xi}[\det(M(\{\xi_1, \xi_2, \ldots, \xi_k, \xi\}, \widehat{\boldsymbol{\theta}}_k))],$$

where k is the number of samples and $\widehat{\boldsymbol{\theta}}_k$ is an estimate based on these samples [107]. Importantly, monotone convergence of this sequence towards D-optimal designs is guaranteed under relatively general conditions [108].
In this section we have focused on design criteria which aim at optimizing parameter estimation. However, alternative criteria have been proposed which target instead model identification [97], model discrimination, hypothesis testing and prediction about future observation, among others [104]. Sometimes combinations of these criteria are used to account for the relative contribution that the experimenter is willing to attach to each of these objectives. Applying these useful criteria in our context is however beyond the scope of the present work.

2.1.2 Dual pulse-chase labeling studies

The standard approach for pulse-chase labeling experiments, including the fraction of labeled mitotic figures (FLM) method, the grain count diminution method and techniques that measure additionally to nucleoside incorporation DNA content, is administration of a single nucleoside analog over a short period of time followed by enumeration of labeled cells or their properties at one or multiple time points after the pulse (see Section 1.2). Over the years, a number of researchers have explored the possibility to use instead of a single pulse, two sequential pulses of nucleoside analogs, in general with the aim to increase the statistical power of the conventional methods.
In 1976 Schultze et al. presented a double-labeling method relying on two radiolabeled nucleosides, namely [^{3}H]- and [^{14}C]- thymidine [22]. Unlabeled, single labeled and double labeled cells could be distinguished with a relatively complex two emulsion autoradiographic technique, and in the framework of the FLM approach a) the duration of S phases, b) cell fluxes at the beginning and end of the S phase, c) cell cycle times and d) sensitivity to drugs (e.g., vincristine) of cells in different phases were determined. Interestingly variances in cell cycle phase completion times and even correlations or more precisely the lack of correlations in subsequent cell cycle phase transitions could also be estimated *in vivo* by this method.
With the halogenated thymidine analog bromodeoxyuridine (BrdU) becoming more popular [109], several studies used a combination of [^{3}H]-thymidine and BrdU labeling [110, 111], where nuclei labeled with [^{3}H]-thymidine were typically detected by autoradiography and nuclei labeled with BrdU by immunocytochemistry. Using this technique, rapid changes in the rate at which terminally differentiating chicken embryo cells entered S-phase could for example be detected with an accuracy not achievable with a single pulse [111]. At the same time, several alternative halogenated thymidine analogs (e.g., iododeoxyuridine (IdUrd), chlorodeoxyuridine (CldUrd)) were discovered and extensively tested [112, 113]. While Pollack et al. described double labeling experiments in mouse mammary tumor grown *in vivo* utilizing IdUrd and CldUrd [33], Ritter et al. [20] assessed the S phase duration (T_S) and the potential doubling time (T_{pot}) of different types of tumors in nine human patients with sequential pulses of BrdU and IdUrd. In this latter study, a simple mathematical method was developed to estimate T_s and T_{pot} from the estimated labeling indices obtained from a single tumor biopsy.
More recently a novel thymidine analog, ethyldeoxyuridine (EdU), has been introduced, which relies on click-chemistry for detection [114, 115]. The main experimental advantage of this analog is the fact that DNA denaturation is not required for proper staining. Because EdU shows little cross-reactivity

with BrdU, double-labeling protocols employing both nucleoside analogs have been developed and applied [116, 117].
While single nucleoside pulse-chase experiments are routinely performed for cell cycle analysis in many laboratories, the double labeling technique, despite its proven statistical power, has found rather limited application. There are probably several reasons for this lack of popularity of dual-pulse labeling among experimentalist. First, the protocol becomes, due to the two labels, more complex, more expensive and more time consuming. Secondly, the analysis and the presentation of the resulting kinetics is not straightforward, and the question what extra information can be gained by a second pulse is not always easy to answer. Finally proper interpretation of the experimental results requires in most cases model-based data analysis.

2.2 Results

2.2.1 Noise explains only part of the parameter identification problem

Measurement noise is the most likely, but not the only candidate reason for failure in parameter identification. Therefore, it is important to test to what extend measurement noise caused the partial identification of the cell cycle parameters when we analyzed proliferating cell cultures in the first chapter of this thesis. If noise would be the sole reason for parameter indetermination, then increasing the number of samples taken at each support point, should lead in a hypothetical limit of infinite sample sizes and under mild conditions, to point estimates for each of the parameters. While it is hard to test such a scenario with real data, it is possible to approximate it, by setting the parameter N in the likelihood function Eq. to increasingly higher values. In the limit of very large N the resulting parameter values are expected to approach the maximum likelihood (ML) point estimates.
In Fig. 2.1 we show for the U87 and the V79 data set the lengths of the 95%-credibility intervals (|CI|) for all six parameters for values of N ranging from $10^2 - 10^7$. One can see for the U87 data set that, while the |CI|s for the G_2M and the S phase vanish with growing N (i.e., $|CI| \propto N^{-1/2}$), the |CI|s corresponding to the G_1 phase converge not to zero but stabilize at around 10 hours. A similar trend is found for the V79 data set, with the difference however that in this case merely the |CI|s for the G_2M phase converge asymptotically to zero.
These observations suggest that even if we had measured the cell cycle kinetics at the original support points a very large number of times, the model parameters could still not be uniquely identified. The simple explanation for this behavior is the fact that many distinct parameter values lead to exactly the same expected measurements at the original support points. In the following section we will therefore change the positions of the support points to better understand the impact of the sampling scheme on asymptotic parameter identification.

2.2.2 Proper sampling is sufficient for parameter identification

As described in Chapter 1, current DNA-BrdU pulse-chase labeling experiments only permit for a limited number of subpopulations (i.e., f^{lu}, f^{ld}, $f^u_{G_1}$, $f^u_{G_2M}$) to be followed separately over time. For instance, labeled undivided cells which are synthesizing DNA are indistinguishable from labeled cells which initiated, but did not complete, the subsequent phase G_2M (together they constitute the f^{lu} subpopulation). Similarly, unlabeled cells in the G_1 phase are, immediately after the pulse, intermixed with unlabeled cells in S phase, unlabeled cell which divided once after label administration and later with second generation progeny of labeled cells (together they constitute the $f^u_{G_1}$ subpopulation). As a consequence, many of the transient population kinetics, which could provide valuable information about the parameters, are hidden during the time window between the pulse and the return to the system's steady state ($f^u_{G_1} = 1$). Therefore, it is not clear whether the information content in the data from conventional DNA-BrdU pulse-chase labeling experiments is sufficient to identify the parameter values, irrespective of the number of samples and the support points we may collect. That the

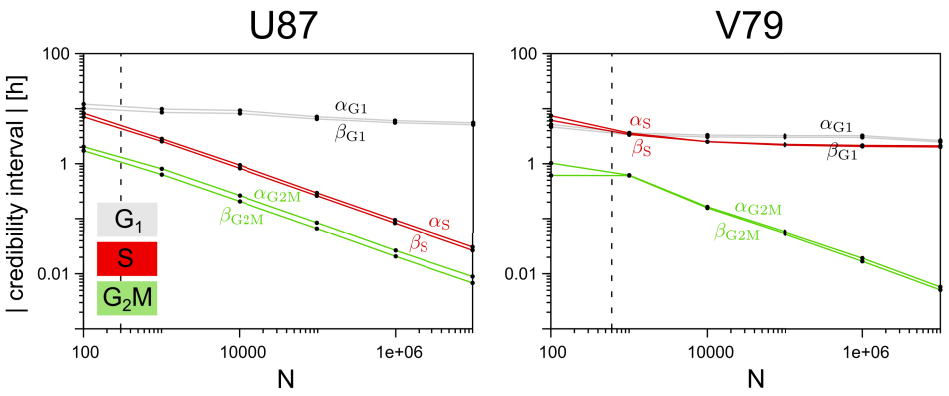

Figure 2.1: Length of the 95%-credibility intervals (|CI|) for cell cycle parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as a function of the sample size parameter N. The |CI|s for the parameters specific for a given phase (same color) are highly correlated and follow the same trends. For the U87 data set the |CI|s decrease according to a power-law $\propto N^{-1/2}$, except for the |CI|s of the G_1 parameters, whose length remains approximately constant at 10 hours. For the V79 data set, the uncertainty in the G_2M parameters drops as in the U87 data set, while the |CI|s for the S and G_1 parameters stabilize after $N > 10000$ at around $2-3$ hours. The dashed lines indicate N estimated from the original data.

information content is indeed sufficient for full parameter identification, at least for the parameters we tested, is the most important finding in this section. Generalizing this result to any parameter set, i.e., by providing a formal proof of uniqueness of the solutions for a given design, appears, due to the memory property of the model and the piecewise-continuous solutions, relatively complex and will not be further pursued here.

Fig. 2.2 shows for simulated data, using ML parameter estimates from the U87 and V79 data sets, approximate bi-variate maximum likelihood regions ($N = 10^6$) for three sampling schemes, differing only in the position of their design points, with the number of support points identical to those used in the true experiments. In the upper row, sampling scheme are chosen such that the support points lie two times closer to the initial pulse; in the middle row the sampling schemes are the one used in the real experiments (U87: $\{0, 2, 4, 6, 8, 12\}$; V79: $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$; in hours, without repeats); and finally at the bottom support points are positioned two times further away from the pulse extending the support points up to 24 and 18 hours for the U87 and V79 data set respectively. The blue circles indicate the parameter values that were used to simulate the data (U87: $\boldsymbol{\alpha} = \{2.3, 3.9, 3.5\}$, $\boldsymbol{\beta} = \{5.7, 4.1, 2.1\}$; V79: $\boldsymbol{\alpha} = \{1.6, 1.1, 0.5\}$, $\boldsymbol{\beta} = \{1.4, 7.8, 1.9\}$). The latter were drawn randomly from the set of maximum likelihood estimates obtained from each data set. In the following, these parameter sets will serve as examples to illustrate the methodology and we will refer to them as $\boldsymbol{\theta}_{\text{U87}}$ and $\boldsymbol{\theta}_{\text{V79}}$ respectively.

For the U87 data set, the approximate ML estimates span from top to bottom, first a 2-dimensional surface, then a line and then a point in the six-dimensional parameter space. The situation for the V79 data set is similar. This example illustrates how the position of the support points dramatically influences, in our model, the information content in the data and determines whether the parameters can be identified or not. Moreover, it shows that the information content of the transient dynamics is in both cases sufficient to identify the parameters, if the support points are well placed. Finally, we can conclude that the support points in the original experiments were apparently chosen too close to the pulse to allow for full parameter identification. Understanding which sampling schemes avoid asymptotic parameter unidentifiability is the task that we want to address in the next section.

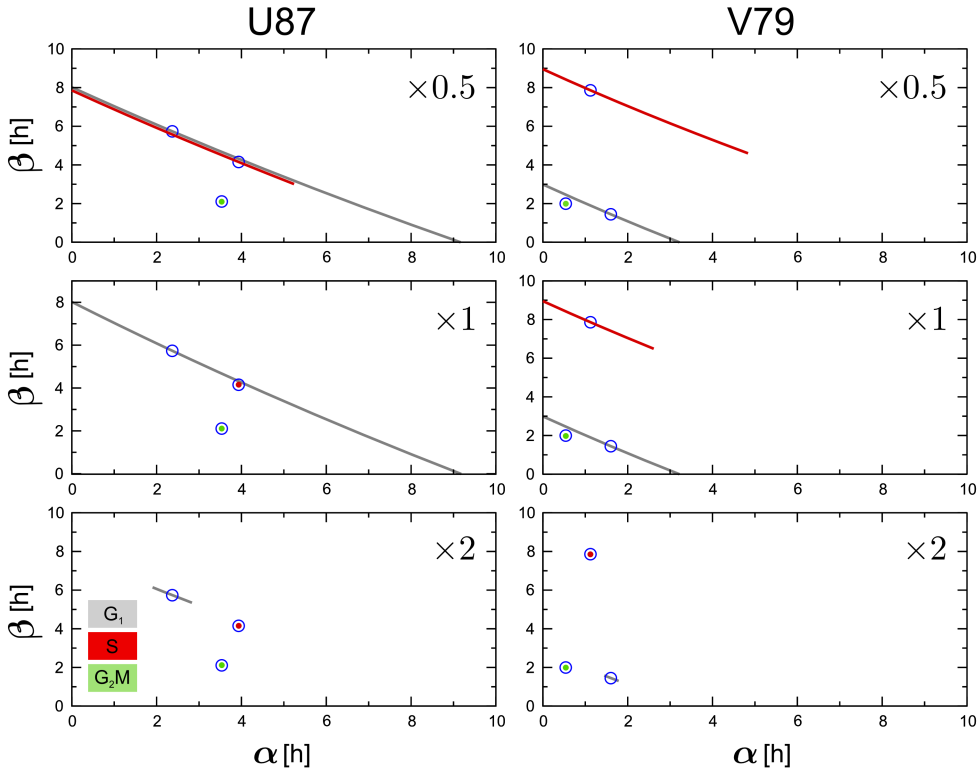

Figure 2.2: Approximate maximum likelihood estimate bi-variate regions (N=10^6) based on simulated data. In the top row the support points were set two times closer to the BrdU pulse compared to the true experiments, in the middle row original sampling schemes were applied (U87: $\{0, 2, 4, 6, 8, 12\}$; V79: $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$; in hours, without repeats), in the bottom row support points were chosen two times further away from the pulse. Parameters used to simulate the data were drawn randomly from the set of maximum likelihood estimates obtained from each data set (U87: $\boldsymbol{\alpha} = \{2.3, 3.9, 3.5\}$, $\boldsymbol{\beta} = \{5.7, 4.1, 2.1\}$; V79: $\boldsymbol{\alpha} = \{1.6, 1.1, 0.5\}$, $\boldsymbol{\beta} = \{1.4, 7.8, 1.9\}$).

2.2.3 Analysis of D-optimal designs for known parameter values

The analytical average kinetics, describing DNA-nucleoside pulse-chase labeling experiments derived in Chapter 1, are piecewise-continuous and depend in a non-linear way on the cell cycle parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Therefore D-optimal designs will also depend, as discussed in Section 2.1.1, on the specific set of parameters. The latter are however in most practical applications not known, to the contrary they are to be estimated. Despite this dependency, which makes it impossible to define a design which is optimal in all situations, it is, as we will see, rewarding to derive D-optimal designs for known parameter sets (e.g., $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$), in order to see whether some pattern or general rules can be identified, that may become useful when parameters are not readily available.

Features of D-optimal sampling schemes

In order to numerically maximize the determinant of the Fisher information matrix, i.e., $\det(\boldsymbol{M})$, we exploit the fact that the likelihood function defined in Chapter 1 (see Section 1.5.2) is everywhere twice differentiable, if the condition $\xi_i \neq \beta_{\mathrm{G_2M}}$ holds for all support points. Excluding this special case from the design space, and under mild conditions, the following expression applies

$$\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}) = -\, E[\boldsymbol{H}(\boldsymbol{z}|\boldsymbol{\xi}, \boldsymbol{\theta})], \tag{2.3}$$

where $\boldsymbol{H}$ is the Hessian matrix of the logarithm of the likelihood function, conditioned on the design $\boldsymbol{\xi}$ and the parameter $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{H}(\boldsymbol{z}|\boldsymbol{\xi}, \boldsymbol{\theta}) = \frac{\partial^2 \log L(\boldsymbol{z}|\boldsymbol{\xi}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}.$$

The expectation operator is averaging over all possible data sets $\boldsymbol{z}$, accounting in this way for all the possible outcomes of an hypothetical experiment that is being designed.
A simplified version of Eq. 2.3 is given by the so-called 'observed' Fisher information matrix, which is evaluated for an 'observed' data set z_0, and consequently the expectation operator can be omitted, i.e., $\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}) = -\boldsymbol{H}(z_0|\boldsymbol{\xi}, \boldsymbol{\theta})$. Here, for computational reasons, we assume that the Hessian matrix evaluated at the most likely data set represents a good approximation for the average $\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta})$. That this assumption is indeed reasonable, was tested by numerically averaging the Fisher information matrix over a large number of simulated data (not shown).
Using the covariance-matrix-adaption evolutionary optimization algorithm [118] and exploiting the high precision and the speed of automatic differentiation [119] for computing the Hessian matrix, D-optimal designs with two and up to six support points were derived. In Fig. 2.3 we show these optimal sampling schemes (filled squares), overlaid for comparison on top of the analytical predictions (solid lines) corresponding to the parameter sets $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$. The sampling schemes used in the original experiments are also provided. Optimal designs, derived following a non-local approach (open squares), including designs with a single support point, will be covered in the next section.
Several features of the D-optimal sampling schemes can readily be identified on the graphs in Fig. 2.3. First, none of the support points is found outside the time window during which the transient kinetics evolve towards the absorbing state (i.e., $\mathrm{f}^{\mathrm{u}}_{\mathrm{G_1}}$=1). This is intuitively clear, because as soon as all cells are in the $\mathrm{f}^{\mathrm{u}}_{\mathrm{G_1}}$ population, nothing except the minimal time until all labeled cells undergo a second division can be inferred. Secondly, support points seem to be placed close to the beginning and towards the end of this time window, with additional ones approximately in the middle and surprisingly none positioned at time zero, immediately after the pulse. Thirdly, designs with more than three support points show repeats (as indicated by the number above the squares) or 'quasi' repeats. This is typically seen for D-optimal designs of linear models, where for uni-variate response functions the number of optimal design points equals the number of parameters. If there are more observation than free parameters, then optimal support points are repeated [120]. Finally, at closer

inspection, the earliest support point is always later than the delay parameter of the G_2M phase. Again this makes sense, because placing the first support point beyond β_{G_2M} allows to extract information from the f^{ld} population, which would otherwise be zero and largely non-informative. Lastly, of note, is the large deviation of the D-optimal designs, when compared with the original sampling schemes. The latter appear to have been 'intuitively' placed such as to cover the kinetics of the f^{lu} population, while the D-optimal designs seem to follow more closely the evolution of the f^{ld} population.

Non-local computation of optimal designs

The fact that one can estimate the credibility regions from the Fisher information matrix at a single point is a formidable simplification, because it allows to deduce from a local property (i.e., a curvature) a non-local property (i.e., a volume). However, this simplification requires that the likelihood function is sufficiently close to a multivariate normal density in order to be valid. For a single design point this condition is apparently not met in our case, as the determinant of the Fisher information matrix yields highly unstable negative and positive values. This is not surprising, as a singular Fisher information matrix indicates local parameter unidentifiability, which is in line with our previous results. Therefore, for our model, the determinant criterion cannot be used to find optimal design with a single support point. For two support points the problem is still not completely resolved, however the determinant is found to be sufficiently well-behaved to permit optimization. And even for three or more support points, some regions in design space still lead to a singular Fisher information matrix.
To derive optimal designs in the singular case or when the normal approximation is not appropriate, we have to adopt a different approach. As we are interested in the sampling scheme that reduces most the volume of the credibility region, we can, in principle, simulate data for each possible design, and then compute the credibility region *via* MCMC. Because in our case, MCMC chains take in the order of 10^4 or more steps to converge, and each step requires an evaluation of the likelihood function, this method is computationally far more expensive than computing the Hessian, which with the help of automatic differentiation approximately requires the time needed for a single evaluation.
Therefore, in order to limit the computational cost, some measures have to be taken which reduce the variability in the volume estimates. These measures consist in fixing the range of possible parameter values from zero to twenty hours and in projecting the 6-dimensional MCMC output onto three 2-dimensional subspaces, where the axis in each of the subspaces corresponds to α and β defining the completion time for each of the cell cycle phases. Then, for each subspace, we compute approximately the 95%-credibility region and sum these surfaces into a final 'volume', which is minimized. For a single support point, Fig. 2.4 shows this quantity over a design space spanned from zero up to forty hours. For one, two and three design points, numerically optimized designs are shown in Fig. 2.3 (open squares). These are in general different from D-optimal design, however approach the latter for three support points. The differences in the volumes computed from D-optimal design and those derived in this section are typically low ($\approx$10%). Surprisingly, in the light of the previous results, for two design points, one of the two measurements lies immediately after the pulse.

Three support points are necessary and sufficient to identify $\boldsymbol{\theta}_{U87}$ and $\boldsymbol{\theta}_{V79}$

To test to what extent the computed optimal designs are more efficient than the sampling schemes from the original experiments, we simulated data and generated, as before approximate maximum likelihood estimates ($N=10^6$) via the MCMC approach. In Fig. 2.5, instead of showing credibility regions, we directly plot the MCMC output for the optimal design with one (direct method), two (D-optimal) and three (D-optimal) support points. The graphs clearly illustrate that a single support point is insufficient to define the parameters. With two support points, a level of 'uncertainty' in the ML estimates is reached that matches closely the uncertainty in the corresponding parameter estimates derived from the original sampling schemes, involving however six and nine support points instead of two for the U87 and the V79 experiment respectively (see Fig. 2.3). Finally with three support points,

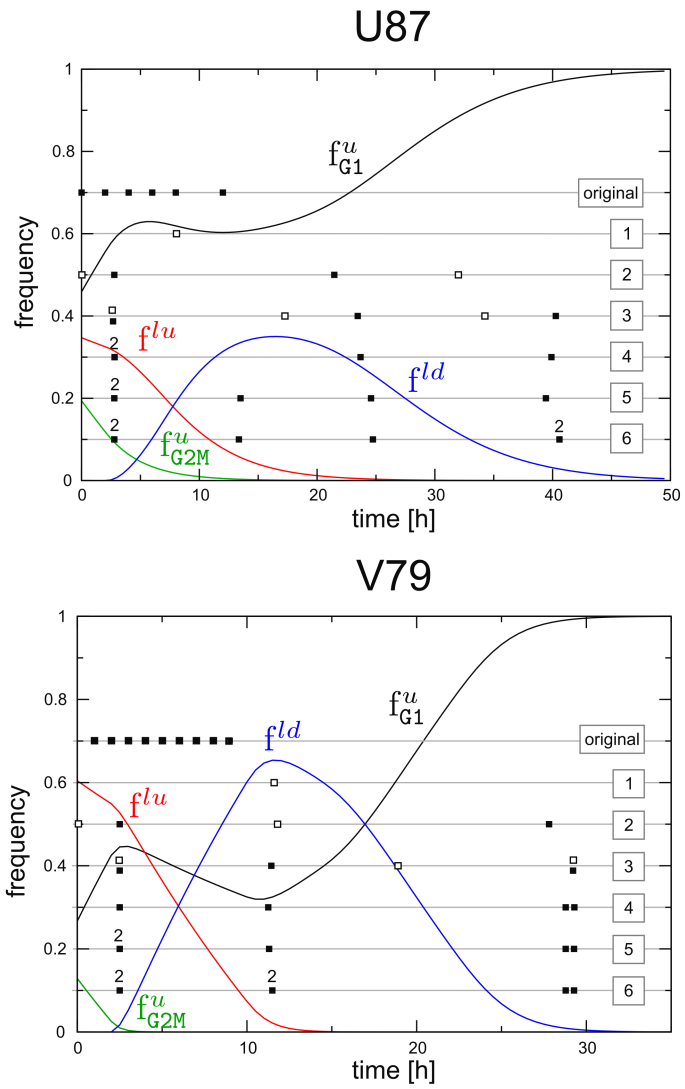

Figure 2.3: Optimal designs for known parameters $\boldsymbol{\theta}_{\text{U87}}$ and $\boldsymbol{\theta}_{\text{V79}}$. D-optimal designs (closed squares) with two and up to six support points, overlaid on top of the model's solutions. Non-local optimal designs (open squares) for one and up to three support points are also shown. They are different from D-optimal designs, but approach these for three support points. For comparison, sampling schemes used in the original experiments are also provided.

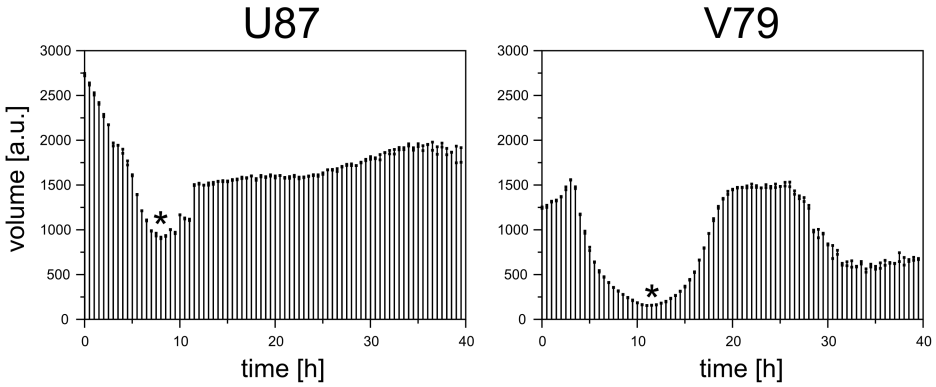

Figure 2.4: Approximate volume (in arbitrary units) of the credibility region for different positions of a single support point. The position which minimizes the volume (star), minimizes the uncertainty in the parameter values, and maximizes the information content of the data.

the estimates approach point estimates and are thus asymptotically identifiable. Together, this shows that three support points are necessary and sufficient to identify the parameter values $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ in the limit of large sample sizes.

2.2.4 Batch-sequential most likely D-optimal design

So far we analyzed D-optimal designs assuming that we knew the underlying parameter values. This of course is not practical, because the parameter values are exactly the quantities that we want to estimate. In this section we will address the more realistic scenario in which the parameters are not known. As outlined in section 2.1.1, batch-sequential design is, under these conditions, a possible strategy to approach the D-optimal sampling scheme in a step-wise fashion. Because we learned before, that three support points seem sufficient to identify the parameters values, we will concentrate on a batch size of three. As a reasonable target, we assume that we want to infer, from a proliferating cell population and under realistic measurement noise, all the model parameter values up to a minimal accuracy of one hour (i.e., $\max(|CI|) \approx 1$ hour). Two questions which become of practical importance are then: how many batches do we need and how many samples should we collect per support point in order to attain this goal? Again we will use the parameter sets $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ and simulated data to exemplify the situation.

Suppose, for the sake of generality, that we don't know anything or almost anything about the cell population under study. We have nevertheless a vague idea about cell cycle kinetics in general. For example, looking at the literature on cell cycle parameter estimates (see Table 2.1), one finds that the average S phase duration lasts between 5 and 27 hours. Or, if one excludes estimates for *in vivo* cancer cells, the average S phase duration seems to lie in between 5 and 11 hours. Bayesian D-optimal design allows to include this kind of knowledge, in form of a prior, into the optimization procedure.

As discussed in Section 2.1.1, one may then decide to maximize Eq.2.2, to derive the most informative design. Unfortunately, this expression has one important limitation, making it unsuitable for our case. It only permits, due to the use of the logarithm, positive values for $\det(\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}))$. It is however common to find, in our situation, for a given experimental design $\boldsymbol{\xi}$, a parameter combination for which $\det(\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}))$ becomes negative, indicating that the likelihood diverges, perhaps only slightly, from the normal distribution. As a more robust alternative, we propose not to choose a design which maximizes the average $\log(\det(\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}))$, but instead to seek after the most likely D-optimal design $\widehat{\boldsymbol{\xi}}$. This can be formalized as follows

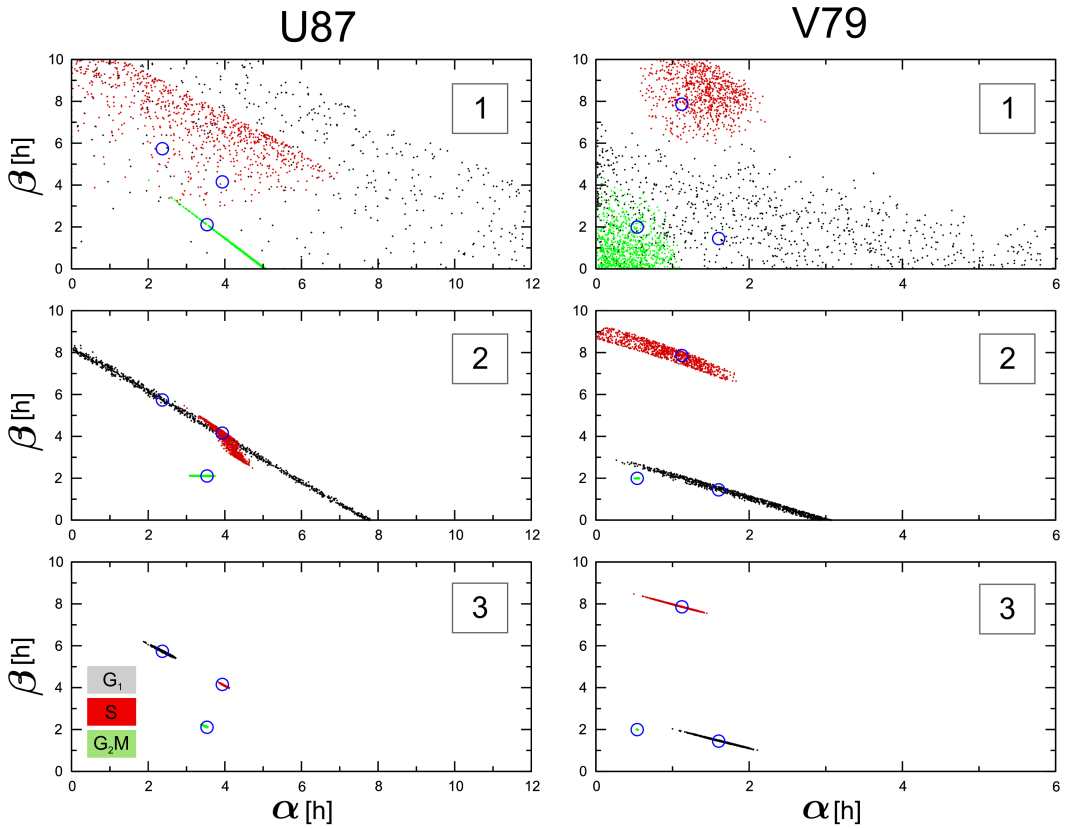

Figure 2.5: MCMC output for approximate ML estimates (N=10^4) showing that three D-optimal support points are necessary and sufficient to identify $\boldsymbol{\theta}_{\text{U87}}$ and $\boldsymbol{\theta}_{\text{V79}}$.

$$\widehat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} \int \delta(\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}))\, p(\boldsymbol{\theta}) d\boldsymbol{\theta},$$

where $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta})$ is the D-optimal design for parameter $\boldsymbol{\theta}$, δ is the Dirac delta function and $p(\boldsymbol{\theta})$ represents the prior knowledge over the model parameters. Thus, by maximizing the integral, which 'enumerates', weighted by the probability of $\boldsymbol{\theta}$, the number of times a given design is D-optimal, we obtain a design which is D-optimal more often than any other design.

A Monte Carlo algorithm to compute the most likely D-optimal design

A Monte Carlo scheme to derive $\widehat{\boldsymbol{\xi}}$ approximately is :

1. draw $\boldsymbol{\theta}_i$ from $p(\boldsymbol{\theta})$,
2. maximize $\det(\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}_i))$ to find $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$,
3. sort the individual design points in $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$ in ascending order,
4. return to 1. until a large number of $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$s are collected, otherwise proceed to 5.,
5. estimate the density in design space from the set of sorted $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$s,
6. search for the global maxima of the density, which should be close to $\widehat{\boldsymbol{\xi}}$.

It is important to notice that the accuracy of the final $\widehat{\boldsymbol{\xi}}$ depends on several factors, first the effectivity of the algorithm that is used to globally maximize $\det(\boldsymbol{M}(\boldsymbol{\xi}, \boldsymbol{\theta}_i))$, then the number of $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$s that are computed to derive the density in the design space, and finally the technique that is used to estimate the density and its maxima from the $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$s. For our analysis, we rely as before on automatic differentiation to compute the Hessian matrix, the CMA evolutionary algorithm is utilized to maximize the determinant and for the density estimation we assume a gaussian mixture model with ten kernels, optimized by the expectation maximization (EM) algorithm (see [121] and Appendix, Algorithm), with a number of $\widehat{\boldsymbol{\xi}}(\boldsymbol{\theta}_i)$s in the range of 500-1000. Finally the global maxima of the density is found by the Broyden-Fletcher-Goldfarb-Shannon (BFGS) gradient search algorithm implemented in the C++ library shark [122] starting from each of the kernels of the Gaussian mixture [123].

Two batches are sufficient to identify $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ asymptotically

For the first batch, sometimes also called the starting design [120], we use values from the literature, excluding estimates from *in vivo* cancer cells, in order to set up the prior. For simplicity, we take an independent normal distribution for the average durations of each phase, with the same mean and variance as the corresponding values in Table 2.1. As variability in phase durations was not reported in most of these studies, we assume that any combination of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that yields the same $\bar{\boldsymbol{\tau}} = \boldsymbol{\alpha} + \boldsymbol{\beta}$ is equally likely. For all subsequent batches we use as a prior the posterior over $\boldsymbol{\theta}$, representing the information from all the data collected in the previous batches.
The initial most likely D-optimal design, given the above specified prior with three support points, is found as $\widehat{\boldsymbol{\xi}}_0 = \{1.2\,\mathrm{h}, 13.5\,\mathrm{h}, 24.8\,\mathrm{h}\}$. This design represents a rational (in contrast to intuitive) starting point for DNA-nucleoside pulse-chase labeling experiments for relatively fast *in vitro* or *in vivo* dividing cells.
We then proceed by simulating a first batch, using $\boldsymbol{\theta}_{\mathrm{U87}}$ as parameter and $\widehat{\boldsymbol{\xi}}_0$ as the sampling scheme, with three repeats and a level of noise mimicking residuals estimated in the original U87 (N≈300) data set. By subsequently examining the artificial data by MCMC, as we did before with the original data, we find that the information content is insufficient to allow for asymptotic parameter identification.

Especially the parameters for the G_1 phase are again poorly resolved (not shown). Nevertheless, when we employ this 'partial' knowledge in form of a prior over the parameters, to derive the next optimal design, we get $\widehat{\boldsymbol{\xi}}_1 = \{3.5\,\text{h}, 24.2\,\text{h}, 43.4\,\text{h}\}$. This design is, despite our ignorance about the parameter values, already very close to the true D-optimal design, which is $\widehat{\boldsymbol{\xi}}_\infty = \{3.4\,\text{h}, 23.4\,\text{h}, 42.4\,\text{h}\}$. In line with previous results, we observe that $\widehat{\boldsymbol{\xi}}_1$ permits full parameter identification in the limit of large sample sizes. Finally for the third batch we obtain $\widehat{\boldsymbol{\xi}}_2 = \{3.2\,\text{h}, 22.9\,\text{h}, 42.6\,\text{h}\}$, confirming that the D-optimal design is, as expected, approaching closer and closer to $\widehat{\boldsymbol{\xi}}_\infty$.
For the parameters $\boldsymbol{\theta}_{\text{V79}}$ the same procedure, with a level of noise estimated from the V79 data set (N≈600), yields $\widehat{\boldsymbol{\xi}}_1 = \{2.5\,\text{h}, 11.6\,\text{h}, 27.4\,\text{h}\}$ and $\widehat{\boldsymbol{\xi}}_2 = \{2.4\,\text{h}, 12.0\,\text{h}, 27.3\,\text{h}\}$ with $\widehat{\boldsymbol{\xi}}_\infty = \{2.2\,\text{h}, 11.9\,\text{h}, 27.5\,\text{h}\}$. Again this indicates that sequential designs converge relatively quickly to the true D-optimal design and allow for full asymptotic parameter identification after about two batches.

A large number of samples are required to infer θ_{U87} and θ_{V79} precisely

Recalling our initial target, namely inferring, under realistic measurement noise, all the model parameter values up to a minimal accuracy of one hour, we can now address the question of how many samples should we collect, say in a third batch, in order to attain this goal. It turns out that, while the parameters for the G_2M are determined with the desired precision using a moderate number of samples, the parameters for the S phase and especially for the G_1 phase are more difficult to infer (see Fig. 2.6). Almost two hundred (for $\boldsymbol{\theta}_{\text{V79}}$) and more than four hundred repeats at the optimal support point are necessary to reduce the uncertainty (95%-credibility interval length) in the S and G_1 parameter up to 1 hour (shown up to 100 repeats in Fig. 2.6). This is far more than most studies can afford, and therefore inferring precise estimates for the variability of cell cycle phases, using this method, seems unrealistic. However important quantities, like the average duration for each phase and the total cell cycle length (see Fig. 2.6, bottom) are estimated with up to one hour precision using less than 10 repeats.
The ability of the method to identify selectively the parameters of the G_2M phase with relative ease, is partly due to the fact that the $\text{f}^u_{\text{G}_2\text{M}}$ population can directly be observed. Moreover, the latter population uniquely depends on the parameters $\alpha_{\text{G}_2\text{M}}$, $\beta_{\text{G}_2\text{M}}$ and μ. In contrast, the three other populations (i.e., f^{lu}, f^{ld}, $\text{f}^u_{\text{G}_1}$) are mixtures and therefore depend in a complex way additionally on all or some of the other cell cycle parameters. This suggests that if we could prevent the intermixing of the different cell populations, perhaps we could reduce the number of repeats that is required to estimate cell cycle parameters with reasonable precision. This is precisely our strategy in the next section, where we explore the potential of dual pulse-chase labeling in order to separate cell populations which become indistinguishable during conventional single pulse-chase labeling experiments.

2.2.5 Dual pulse-chase labeling improves the quality of parameter estimates

As discussed in Section 2.1.2, dual pulse-chase labeling (DPL) experiments have been carried out in the past, in general with the aim to increase the statistical power of the conventional single pulse-chase labeling (SPL) experiments. Simple theoretical models have been developed in some of these studies, to explain and interpret the kinetics obtained from DPL experiments, where the focus lay for example either on the total number of single-labeled and double-labeled cells without distinction of the cell cycle phase [20], or in the context of the FLM method on the fraction of single-labeled and double-labeled cells in mitosis [22]. The combination of DPL with DNA content staining (e.g., DAPI), which provides additional information about the cell cycle phases, remains however to our knowledge experimentally and especially theoretically, largely unexplored.
The results in this section will confirm previous insights on the statistical power of DPL. The model developed in Chapter 1 is employed to describe unlabeled, labeled and double-labeled cells in each of the three cell cycle phases G_1, S and G_2M. D-optimal designs are derived and their increased efficiency is compared to D-optimal SPL experiments. Finally, in line with the last point in the previous section,

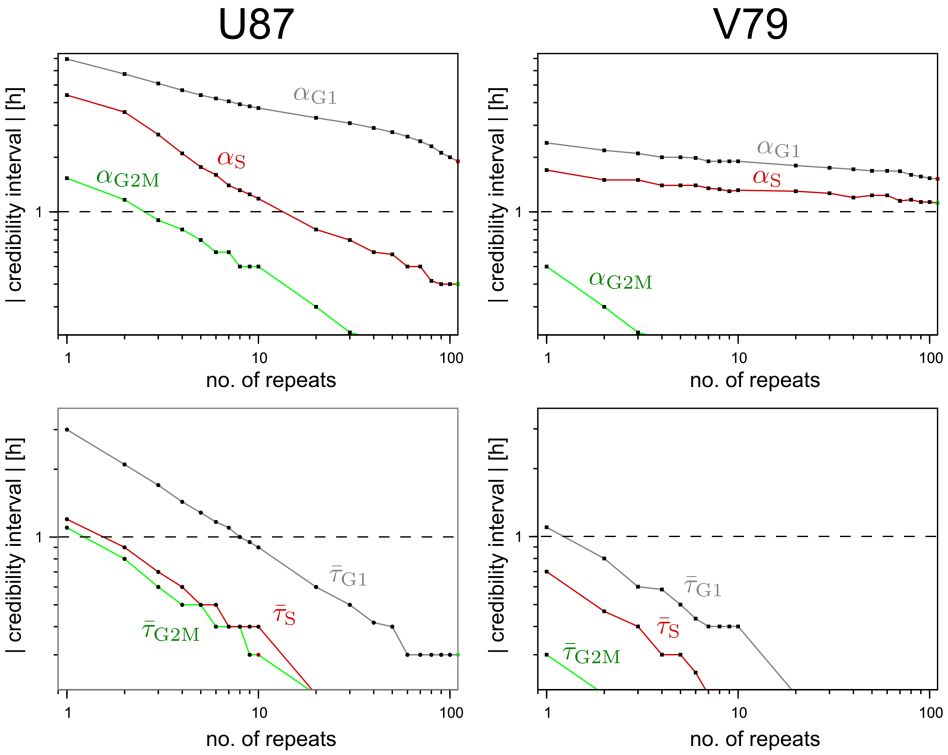

Figure 2.6: 95%-credibility interval lengths (|CI|) for the standard deviations (top row) and the average durations (bottom row) of the phase completion times as a function of repeats at three approximate D-optimal support points, derived from batch-sequential design. The |CI|s for the delays (β_i) are very similar to the standard deviations (α_i) and are therefore not shown. One can see that while the average duration for each phase are estimated with reasonable precision using about ten repeats, the uncertainty in α_{G_1} remains, even after one hundred repeats, above the target of one hour (dashed line).

τ_{G_1}	τ_S	τ_{G_2M}	total	cell line	*in vivo* cancer	reference
5.12	5.02	1.76		WB344		[124]
	9.6			647V	x	[20]
	10.9			MC26	x	id
	7.7			colorectal	x	id
	27.2			colorectal	x	id
	14.8			colorectal	x	id
	18.2			colorectal	x	id
	6.8			ret. trigone	x	id
	23.0			cervix	x	id
	10.8			floor of mouth	x	id
	13.2			recur. prostate	x	id
	16.6			renal	x	id
3.7	7.8	1.8	13.3	murine jenumum		[10]
9.5	7.5	1.45	18.7	id		id
4.5	5		11.5	id		id
4.7	6.7	1.5	13.1	id		id
3.2	8.3	1.65	13.1	id		id
3.6	7.6		12.42	id		id
4.5	8.0	1.5	14.0	id		id
3.7	7.8		13.3	id		id
5.6	7.0	1.4	14	CHO		[18]
4.6	6.3	1.7	12.6	CHO		id
	10	3.6		murine lung		[17]
39.4	9.4	11.7	60	Kag cancer	x	[21]
13.8	14.8	9.7	38.3	id	x	id
13.4	13.1	19.6	46.1	Vaa cancer	x	id
20.9	10.4	14.5	45.8	id	x	id
34.9	6.3	21.9	63.1	ScuL cancer	x	id
6.2	9.9	3.1	19.2	ScuS cancer	x	id
6.6	8.0	5.9	20.5	id	x	id
31	17.5	4.6	53.1	Fro cancer	x	id
10	12.4	5.4	27.8	id	x	id

Table 2.1: List of cell cycle parameter estimates from the literature (all values in hours).

batch-sequential design is carried out *in silico*, and the number of batches and the number of repeats necessary for reasonable precision and accuracy in the parameter estimates is assessed. Experimental validation of the results, although essential, is out of the scope of the present investigation, and is scheduled for future work (see Fig. 2 for a preliminary experiment in which Jurkat cells were pulse-labeled first with EdU, then with BrdU).

Nine instead of four subpopulations identifiable by DNA-nucleoside DPL

While there exist many possible variants of DPL (see Section 2.1.2), the protocol we consider here, uses two different nucleoside analogs which, when incorporated into the DNA of a cell, are specifically detectable concurrently with DNA content by FACS analysis. A candidate pair of analogs could be, for example, the commonly used BrdU together with the recently discovered thymidine analog EdU, however the method is not restricted to these.

An important characteristic of the protocol, which distinguishes it from those used in the studies discussed in Section 2.1.2, is that one of the two pulses is always placed immediately before cell fixation (see Fig. 2.7). The latter strategy bears several conceptual, theoretical and experimental advantages. First, from an experimental point of view, this protocol, if applied in an *in vivo* context, allows to label cells *ex sito* with the second nucleoside analog. As a consequence, the amount of possibly carcinogen nucleoside analog the organism is exposed to is reduced. Secondly, the populations that can be identified by DPL in combination with measuring the DNA content, are, as we will see, easy to interpret, and represent cell populations with are, in contrast to a protocol in which the second pulse is not administrated immediately before fixation, 'pure' and not mixed in respect to their cell cycle phase. Thirdly, mathematical treatment is significantly simplified, especially because predictions for double-labeled cells are much more complex than their single-labeled counterparts. Finally, single-pulse optimal designs with two support points, derived with the non-local method in Section 2.2.3, were found to have one of the two support points placed at time zero. Because dual pulse labeling data contains (i) the same information than two single pulse labeling experiments plus (ii) the extra information from the double-labeled population, the results from optimal single pulse labeling indicate that the information in (i) may be maximized by the protocol. We will come back to this point in the discussion.

To appreciate the additional populations identified by DPL, data from a single pulse-chase labeling experiment was artificially colored, to mimic the expected FACS output from proliferating cells labeled according to the protocol described before. In Fig. 2.8, besides the usual gates defining the populations f^{lu}, f^{ld}, $\mathrm{f}^{u}_{\mathrm{G_1}}$ and $\mathrm{f}^{u}_{\mathrm{G_2M}}$ as before (see Fig. 1.3), cells that have incorporated the second label are drawn in red and subpopulations that can be identified based on the two labels are indicated by rounded squares. For the time immediately after the pulse (i.e., $t = 0$), no extra information is gained by the second pulse. However, already two hours later, two additional population can be discerned. Twelve hours after the first pulse, seven population, instead of three, can be recognized. Thus by resolving each of the four initial population according to the cell cycle phases, it is possible to measure the kinetics of in total nine populations ($\mathrm{f}^{lu} \to \{\mathrm{f}^{lu}_{\mathrm{S}}, \mathrm{f}^{lu}_{\mathrm{G_2M}}\}$, $\mathrm{f}^{ld} \to \{\mathrm{f}^{ld}_{\mathrm{G_1}}, \mathrm{f}^{ld}_{\mathrm{S}}, \mathrm{f}^{ld}_{\mathrm{G_2M}}\}$, $\mathrm{f}^{u}_{\mathrm{G_1}} \to \{\mathrm{f}^{u}_{\mathrm{G1,G1}}, \mathrm{f}^{u}_{\mathrm{G1,S}}, \mathrm{f}^{u}_{\mathrm{G1,G2M}}\}$ and $\mathrm{f}^{u}_{\mathrm{G_2M}} \to \{\mathrm{f}^{u}_{\mathrm{G_2M,G_2M}}\}$). Because all these kinetics depend on the cell cycle parameters, each of them can in principle tell us something about the phase completions times. However some information is redundant. For example if $\mathrm{f}^{lu}_{\mathrm{S}}$ and $\mathrm{f}^{ld}_{\mathrm{S}}$ is measured, then $\mathrm{f}^{u}_{\mathrm{G1,S}}$ is defined by the total fraction of cells in S phase, because $n_{\mathrm{S}} = \mathrm{f}^{lu}_{\mathrm{S}} + \mathrm{f}^{lu}_{\mathrm{S}} + \mathrm{f}^{u}_{\mathrm{G1,S}}$. Similarly from $\mathrm{f}^{lu}_{\mathrm{G_2M}}$, $\mathrm{f}^{ld}_{\mathrm{G_2M}}$ one can deduce $\mathrm{f}^{u}_{\mathrm{G_1,G_2M}} + \mathrm{f}^{u}_{\mathrm{G_2M,G_2M}}$, by knowing the frequency of cells in $\mathrm{G_2M}$ phase.

Notice that the predictions for all 'new' populations are readily given by the solutions Eq. 1.26 derived in Chapter 1 (e.g., $\mathrm{f}^{lu}_{\mathrm{S}} = n^{\mathrm{S}}_0$; $\mathrm{f}^{lu}_{\mathrm{G_2M}} = n^{\mathrm{S}}_1$; $\mathrm{f}^{ld}_{\mathrm{G_1}} = n^{\mathrm{S}}_2$; $\mathrm{f}^{u}_{\mathrm{G_2M,G_2M}} = n^{\mathrm{G_2M}}_0$). In contrast to the treatment of SPL experiments, pooling together several populations has become largely unnecessary. Thus, interestingly, a more complex experimental design leads, in this case, to a more simple mathematical treatment.

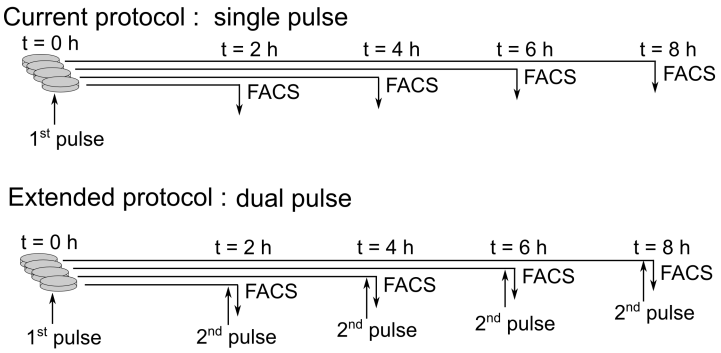

Figure 2.7: Single and dual pulse-chase labeling protocol for cell culture experiments.

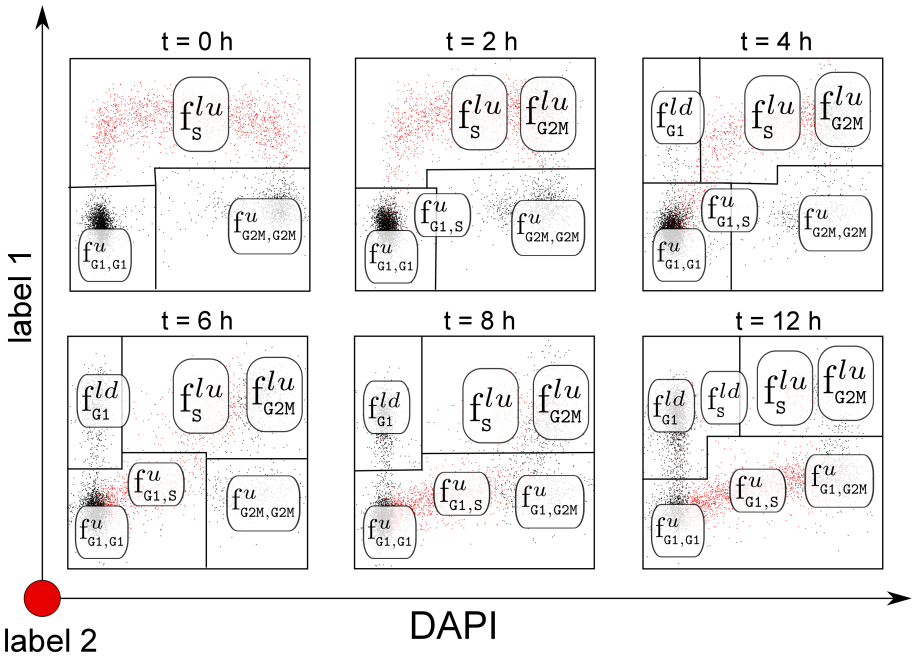

Figure 2.8: Artificial staining of SPL data, showing seven of the nine subpopulations that could potentially be identified with DPL. Notice that the four population f^{lu}, f^{ld}, $f^{u}_{G_1}$ and $f^{u}_{G_2M}$ that can be followed by SPL, have each been subdivided according to the cell cycle phases.

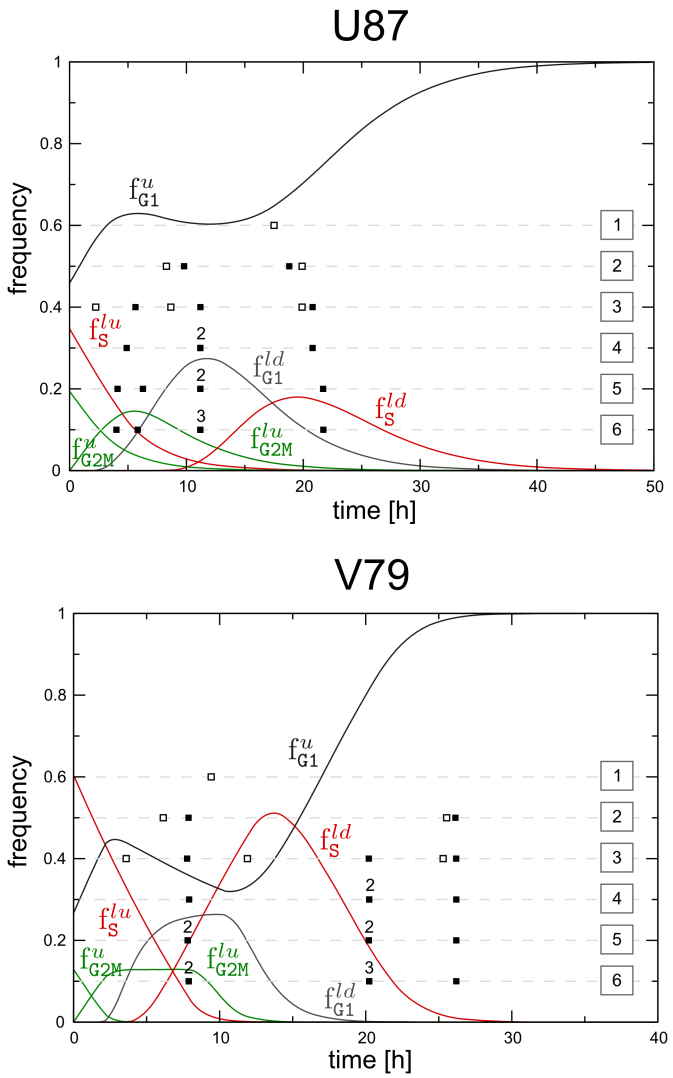

Figure 2.9: Optimal designs for known parameters $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$. D-optimal designs (closed squares) with two and up to six support points, overlaid on top of the DPL model solutions. Non-local optimal designs (open squares) for one and up to three support points are also shown.

Two support points are necessary and sufficient to identify $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$

Following mostly the steps in the previous section, we first analyze optimal designs for known parameter values. For $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$, Fig. 2.9 shows D-optimal design (filled squares) and non-local optimal designs (open squares) for the DPL protocol. As a general trend, the optimal design points for DPL differ from those for SPL experiments and are positioned closer to time zero. As before, however, design points are placed within the time window between the last pulse and the return to the system's absorbing state (i.e., $\mathrm{f}^{\mathrm{u}}_{\mathrm{G_1}}=1$). When testing the optimal design with one and two support points in the limit of large sample sizes, it turns out that two but not one support points are sufficient to reach asymptotic parameter identifiability (not shown).

Batch-sequential designs for $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ converge approximately after two batches

As before, we initiate the batch-sequential design procedure with the most likely D-optimal design, based on a prior reflecting literature data (Table 2.1 excluding *in vivo* cancer studies). We obtain for

the first batch $\widehat{\boldsymbol{\xi}}_0 = \{1.2\,\mathrm{h}, 5.3\,\mathrm{h}, 12.2\,\mathrm{h}\}$, which measures the kinetics at somewhat earlier times than the starting design for SPL experiments.
For $\boldsymbol{\theta}_{\mathrm{U87}}$, using simulated data with three repeats and appropriate noise, we then find for the second optimal batch $\widehat{\boldsymbol{\xi}}_1 = \{6.2\,\mathrm{h}, 11.3\,\mathrm{h}, 21.2\,\mathrm{h}\}$. In the case of $\boldsymbol{\theta}_{\mathrm{V79}}$, we compute the most likely D-optimal second batch as $\widehat{\boldsymbol{\xi}}_1 = \{8.0\,\mathrm{h}, 20.4\,\mathrm{h}, 25.4\,\mathrm{h}\}$. These second optimal batches indicate that the designs converge, as it was the case for SPL, relatively quickly towards the true D-optimal designs, with are derived for DPL as $\widehat{\boldsymbol{\xi}}_\infty = \{5.64\,\mathrm{h}, 11.2\,\mathrm{h}, 20.8\,\mathrm{h}\}$ and $\widehat{\boldsymbol{\xi}}_\infty = \{7.8\,\mathrm{h}, 20.2\,\mathrm{h}, 26.1\,\mathrm{h}\}$ for $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ respectively.

Around ten repeats are required to infer $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ precisely

With a good approximation of the D-optimal design in hands, which should guarantee, as we have seen before, parameter identification, we are now in the position to examine the important question of how many samples we need to estimate the parameters precisely. As done in the corresponding section on SPL, we compute the 95%-|CI|s for increasing number of repeats. Furthermore the parameter N, which specifies the effective population size, was fixed as before to 300 and 600 for the parameters $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$ respectively.
In Fig. 2.10 we show, using the optimal designs from the second batch, the 95%-|CI|s of the α parameters (top row), corresponding to the standard deviations of the phase completion times, and the 95%-|CI|s for the average duration of each phase $\overline{\tau}_{\mathrm{G_1}}$, $\overline{\tau}_{\mathrm{S}}$ and $\overline{\tau}_{\mathrm{G_2M}}$ (bottom row). One can see that for 10 repeats, all parameters are determined up to one hour precision. This is about forty times (!) less than the number of repeats required for the same parameter values, using SPL instead of DPL. Being able to resolve each of the four population from SPL according to their cell cycle phases thus comes with an enormous increase in precision.
Finally it is noticed that the least determined parameter always dominates the estimates for the total cell cycle. For example, if we want to estimate the variance in the cell cycle duration, i.e., $\sigma_T^2 = \sigma_{\mathrm{G_1}}^2 + \sigma_{\mathrm{S}}^2 + \sigma_{\mathrm{G_2M}}^2$, then even if we know σ_{S} and $\sigma_{\mathrm{G_2M}}$ exactly, uncertainty in $\alpha_{\mathrm{G_1}} = \sigma_{\mathrm{G_1}}$will prevent precise inference of σ_T^2. Therefore, even if only the average or variability in the cell cycle duration is of interest, it is important, with this method, to reduce uncertainty in the corresponding quantities for all cell cycle phases.

2.3 Discussion

In this chapter, we have analyzed ways to improve DNA-nucleoside pulse-chase labeling experiments. First, we optimized the sampling schemes of single pulse-chase labeling experiments for known parameters, and found that three D-optimal support point are necessary and sufficient to determine parameter values in the limit of large sample sizes. Then we used batch-sequential design to derive most likely D-optimal support points for the more useful case in which the parameters are not known. We found that two or three batches were sufficient to approach very closely the D-optimal design, but had to realize that far too many repeats are required to infer all the parameters with reasonable precision. To overcome this limitation, we modified, as previously done by others, the conventional single-pulse method, and added a second nucleoside analog pulse to the protocol. This reduced *in silico* significantly the number of repeats required for high quality parameter estimates, suggesting DPL in combination with the model-based inference strategy developed herein as a promising new method to infer cell cycle parameters with high precision. There remained however some aspects of the methodology which could not fully be explored.
One obvious flaw of the work presented herein, is the fact that the results have not been verified by 'real' experiments. Attempts to implement the DPL protocol have unfortunately remained preliminary (see a preliminary experiment in Fig. 2 where Jurkat cells were pulse-labeled first with EdU, then 30 minutes later with BrdU). Nevertheless, even though the results are so far purely based on simulations,

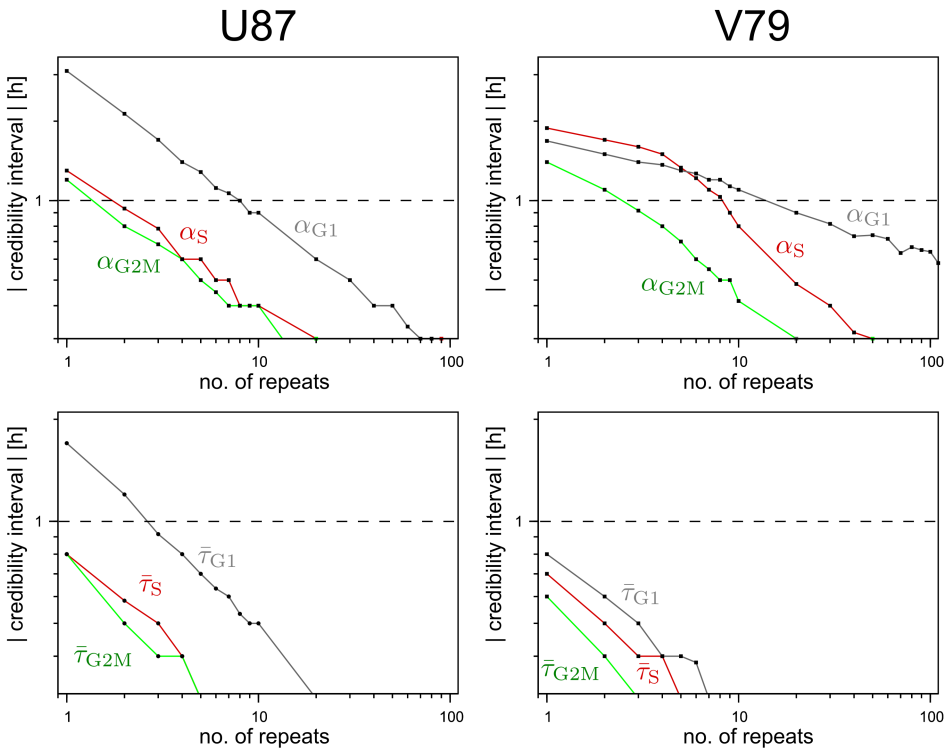

Figure 2.10: 95%-credibility interval lengths (95%-|CI|) for the standard deviations (top row) and the average durations (bottom row) of the phase completion times as a function of repeats at $\widehat{\boldsymbol{\xi}}_1$ for DPL experiments. This shows that, for the same number of repeats, DPL yields more precise estimates for all cell cycle parameters when compared to SPL (see Fig. 2.6). Moreover, ten instead of four hundred repeats, are sufficient to reach our target of a minimal 95%-|CI| of one hour.

we are confident that they will hold in the real setting. Out of prudence, one might however ask, which 'surprises' are to be expected, when applying the proposed methodology in the laboratory. First of all, the model might not reproduce as well the kinetics obtained from DPL than those from SPL. If this is the case, the model can be modified, for example, by adopting a delayed hypoexponential completion time for each phase, instead of the more simple delayed exponential density. The theory, especially the way we solve the average kinetics and the way we derive D-optimal designs, would essentially remain the same. Another complication that may arise are the unknown long-term (>24 hours) effects of the nucleoside analogs BrdU and EdU on the proliferative behavior of labeled cells. However, this issue is common to all nucleoside pulse-chase experiments [125], and minimizing the amount of label incorporation should help to reduce the impact of this potentially harmful factors.
As mentioned before, placing the second pulse in DPL immediately before cell fixation has a number of advantages. However, we could not conclusively show that this represents the best design strategy. To confirm this, one would need to derive predictions for double-labeled cells, which is not straightforward due to their complex history which has to be taken into account. Even though, in Chapter 3, we will solve kinetics of double-labeled cells, we only consider double-labeled cells which are and remained in S phase during the two pulses. This is justifiable in that case, because the time lapse between the two pulses, specified by the original experiments, was only three hours.
Finally, after analyzing the performance gain of DPL over SPL, one might naively ask whether triple pulse-chase labeling (TPL) could further improve the potential of kinetic cell cycle experiments. The answer is : Yes. Recalling that DPL contains as a lower bound the information of two SPL experiments, TPL contains at least the information of three SPL experiments. Because we could show that three D-optimal SPL support points are sufficient to asymptotically identify the parameters $\boldsymbol{\theta}_{\mathrm{U87}}$ and $\boldsymbol{\theta}_{\mathrm{V79}}$, we can anticipate, that a single optimal TPL sample will be sufficient for parameter identification in the limit of large sample sizes in this case. Moreover, with the additional information of double- and triple-labeled cells the method will certainly outperform SPL and DPL in terms of statistical power. And the fact that a single sample might be sufficient for complete parameter identification, represent an interesting perspective for *in vivo* experiments or even for potential clinical applications, where taking a biopsy from the tissue of interest (e.g., a tumor) more than a single time might not be possible. While double pulse labeling (without DNA content staining) have been studied since the late 80s, triple labeling with thymidine analogs is a technique that has not been envisaged until recently [126, 127].

3 Proliferation and Migration in Germinal Centers

3.1 Motivation and Background

In the first two chapters of this thesis, our focus lay on estimating cell cycle parameters from exponentially growing cell populations. While arguably of practical importance, for instance in the analysis of cell culture growth, initial stages of immune responses and acute lymphomas, unlimited or free growth, due to its transient nature, represents necessarily a rather exceptional scenario. More common, and therefore probably more relevant, are situations of cell populations proliferating under homeostatic control, wherein the number of newly produced cells is counter-balanced by cell death, emigration or differentiation. Examples of homeostatically controlled cell population include most if not all healthy tissues in human adults, stem cells, naive immune cells, red blood cells, chronic lymphomas and other temporarily non-growing cancers.
In this chapter, we will expand the concepts and the modeling approach, proven successful before in the *in vitro* context, to a more complex *in vivo* environment, presumably under temporary homeostatic control. The specific environment we will focus on are transient anatomical structures known as germinal centers (GC), which play a key role in the humoral arm of the adaptive immune response. Class switch recombination, somatic hypermutation, affinity maturation as well as clonal expansion and differentiation into plasma and memory B cell all occur or at least depend on GCs (for more details see Section 3.1.1).
As was shown by many years of research, central for performing these distinct, yet interleaved functions, is cell division. Studying cell cycle kinetics in GCs is however hindered by the fact that the latter are heterogeneous dynamic entities, involving a complex interplay between proliferation, cell migration, differentiation and affinity selection. In this chapter, by integrating some of these processes into a model-based cell cycle analysis, we aim not only at improving our understanding of proliferation in GCs but also at reconciling some of the apparent contradictory findings about GCs that have been reported in the literature.
After a short introduction into the humoral immune response, the role of GCs and their spatial organization are briefly reviewed. Because GCs represent an especially fertile field in theoretical immunology, the major insights gained from GC modeling studies are recapitulated. Finally, former studies which reported on GC cell cycle estimates are reviewed. In the results section, two different stochastic models are developed, solved and compared to five previously published data sets. With the first model, overall proliferation and selection in the GC B cell population is analyzed, using *in vivo* DAPI-BrdU pulse-chase labeling data of GC B cell populations. With the second model, proliferation and migration is studied, specifically in the so-called dark zone (DZ) of GCs, relying on published data from BrdU double-pulse labeling, *in vivo* photoactivation and stathmokinetic experiments. The main findings from this modeling attempt, namely the existence of a transiting subpopulation in the DZ and the influx of cycling cells from outside into the DZ of GCs are elaborated on in the discussion.

3.1.1 Characteristics of germinal centers

According to current beliefs, a typical humoral immune response is initiated when one or several recirculating naive B cells bind *via* their surface B cell receptors with sufficient 'strength' to antigen-antibody immune complexes on follicular dendritic cells (FDC) in primary follicles of secondary lymphoid organs [128]. Activated B cells then internalize the antigen and migrate to the periphery of the follicle in which they became activated [129]. Here they seek to interact and receive help from

activated T helper cells by antigen-derived peptide-MHC (pMHC) complexes presentation on their surface, in order to become fully committed [130].
Antigen-specific T helper cells, on their turn, are activated by dendritic cells in the T cell zone, however only if the latter had previously internalized the antigen or antigen-antibody immune complex by unspecific phagocytosis [131] and are thus capacitated to present, like the activated B cells, antigen-derived pMHCs on their surface.
After full activation, B cells then either commit to the so-called extra-follicular B cell response or engage into the formation of GCs. In the first case, the activated B cells undergo rapid clonal expansion followed by differentiation into short-lived plasma cells. Plasma cells are able to produce quickly large amounts of low-affinity germline encoded antibodies or somatically hypermutated antibodies, if specific memory B cells from a previous encounter with the antigen become activated [132]. B cells that engage into the formation of GCs return to the follicle in order to found, with the help of cognate follicular T helper cells, FDCs, tingible body macrophages and probably a small number (e.g., 5-50) of other clonaly unrelated but antigen specific B cells a new germinal center [133, 134].
Typically, a few hundred such new germinal centers are formed in secondary lymphoid organs over the course of several weeks after antigen challenge. Depending on the experimental protocol, peak GC B cell numbers are observed between 5-10 days after immunization, which slowly wane away in the subsequent weeks [134, 135]. The processes associated with the appearance and the disappearance of GCs have collectively been termed the germinal center reaction (GCR). Its fundamental role in adaptive humoral immunity has extensively been studied and will be described in more detail in the following sections.

The role of germinal centers

The role of GCs in adaptive humoral immunity is at least three-fold. The GCs provide microenvironments or niches within which (i) affinity maturation, (ii) differentiation into high-affinity plasma and memory cells and (iii) class switch recombination occur [135].

- **Affinity maturation** denotes the increase in average affinity of serum antibodies towards the epitopes of a given antigen after immunization. Initially discovered by Eisen *et* al. [136], it has later been postulated that affinity maturation is driven by a GC restricted Darvinian-like process of clonal expansion, mutation and selection, favoring survival of B cell clones bearing hypermutated surface antibodies with higher affinity towards prevailing antigens [137, 138]. This view is supported by several classical and recent experimental observations which show a remarkable combination of proliferation, mutation and selection sharply localized to GCs. First, proliferative activity in GCs is significantly higher when compared to other locations in lymphoid organs or the blood [139, 140]. This indicates that GCs are the major site of B cell clonal expansion during adaptive immune responses. Secondly, somatic hypermutation, i.e., random point mutations mostly introduced into the IgV region of B cell receptor genes, and the associated expression of activation induced cytidine deaminase (AID), was also found highly restricted to GCs [141]. Thirdly, selection occurring in GCs has been 'demonstrated' by sequencing B cell receptor encoding genes from e.g., single cells picked from serial section of GCs [133, 142–144]. Genealogies generated from these trees suggest an ongoing selection process, with a high ratio of replacement versus silent mutations and affinity enhancing mutation acquired in a step-wise fashion [145]. However the interpretation of these findings is still a matter of debate, mainly due to the limited number of receptors sequenced, an issue that might hopefully be resolved in the near future by next generation sequencing.

- **Differentiation** of GC B cells into memory B cells and long-lived plasma cells is known to be initiated within GCs, even though the exact mechanism remains to be elucidated [146]. Both cell phenotypes, memory B cells and long-lived plasma cells are considered as the final outcome of an adaptive humoral immune response and together are complementary in providing protective

humoral immunity. After encounter with cognate antigen, memory B cells undergo more rapid reactivation and mount a faster and more potent humoral immune response than their naive inexperienced counterparts [147]. Plasma cells, on their turn, are highly hypermutated and are found mostly in the bone marrow, a niche which is known to promote their survival. Here they produce, over extended periods of time, antibodies which diffuse into the body fluids, where they mediate clearance of any object they bind to [148]. As mentioned before, little is known about the mechanism that lead inside or outside GCs to one or the other fate [149], but it has been suggested that high affinity GC B cell preferentially become plasma cells [150].

- **Class switch recombination** (CSR) is, as somatic hypermutation, an AID dependent mechanism [151], and as such probably largely restricted to GCs. CSR allows plasma cells to target their effector function (through diffusing immunoglobulin) to specific anatomical sites in the body, by switching the gene fraction coding for the constant regions of the immunoglobulin heavy chain., i.e., the class (IgA, IgE, or IgG) of the antibodies, they produce. Besides of determining its preferential location in the body, the class of a given antibody also determines its potential life time and influences the behavior of other components of the immune system, e.g., the complement system and macrophages, it interacts with. Despite the important role in 'focusing' the humoral immune response to certain environments in the body, the way by which GC B cells 'decide' to which class they switch remains, as with the fate decisions regarding plasma and memory differentiation, enigmatic [147].

Germinal center organization

In order to carry out their distinct functions, GCs seem to be tightly organized both spatially and temporally [135]. A spatial compartmentalization, that can be readily visualized by immunohistochemistry based on staining pattern, are the light zone (LZ) and the dark zone (DZ). Historically, B cells in the LZ are called centrocytes while B cells in the DZ are termed centroblasts.

- The **LZ** harbors most of the follicular dendritic cell network and attached iccosomes (immune complexes in highly ordered units). It also shows a higher density of follicular T helper cells [152], which are known to provide important pro-survival signal to GC B cells *via* direct cell-cell contact and cytokine secretion. Together these observations have been suggestive for a model, in which selection of GC B cells bearing antibodies with higher affinity towards the prevailing antigen is mainly carried out in the LZ, however no definitive consent has been reached [26].

- Cells in the **DZ** show higher expression of AID [153], indicating that somatic hypermutation and class switch recombination preferentially occur in this zone. There is also a clear bias of mitotic figures in the DZ when compared with the LZ [36]. This fact had been interpreted in the past as an indication that cell proliferation is restricted to the DZ. BrdU pulse labeling [24, 37] and Ki-67 staining [135] have however repeatedly shown that cells in both zones are actively cycling. Reconciling both observations, recent experimental work indicates that not proliferation but cytokinesis exclusively occurs in the DZ [153].

Over the years a general picture emerged, based on live-imaging, histoimmunological staining and gene expression studies, in which GC B cells in the DZ and LZ represent relatively similar phenotypes, with only minor but reproducible differences in gene expression profiles [26]. It has been proposed that GC B cells may switch rapidly from one phenotype to the other in a possibly cell cycle specific manner [141].

Germinal center migration

Although spatially segregated, the DZ and LZ cell populations are not completely isolated from each other. Cell migration from the DZ to LZ, initially proposed to explain [^{3}H]-thymidine pulse labeling

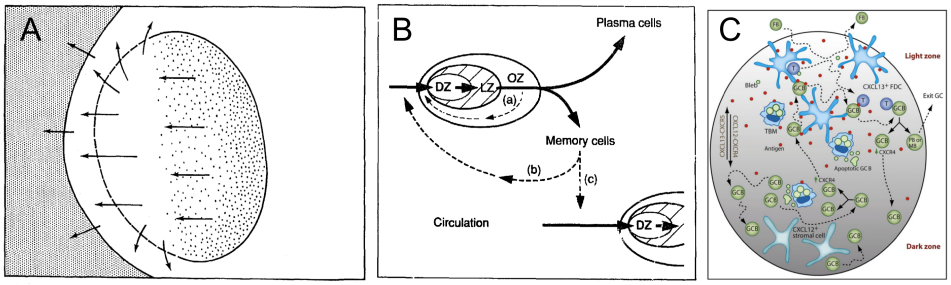

Figure 3.1: Evolution of germinal center models over the last fifty years. Graphs are reproduced from the original papers for historical reasons (A, [61]; B, [154]; C, [26]).

experiments in the early sixties [61], was confirmed in 2007 directly by two-photon imaging in lymph nodes of living anesthetized mice [24, 25, 37]. Migration in the opposite direction, i.e., from the LZ to the DZ, postulated by mathematical models of affinity maturation [154], was also observed in these studies. The reported quantitative estimates of migration rates between the zones remained however controversial, mainly due to the relatively short recording times and the limited volume that could be imaged. This issue was resolved to a certain extend in 2010 by a novel experimental approach in which GC B cells in the DZ and the LZ of genetically modified mouse were photoactivated and tracked for up to six hours [153]. This confirmed again cell migration in both directions with a clear net flux from the DZ towards the LZ. Unfortunately these technically challenging experiments have not been reproduced or published yet by other research groups.

3.1.2 Cell cycle analysis in germinal centers

Many different methods have been employed to infer GC cell cycle parameters in both mice and rats (see Table 3.1). Early studies by Hanna *et* al. in 1964 [61] made use of radioactive thymidine which gets selectively incorporated into the DNA of cells that are in the process of replicating their genome. By following cells that became labeled by a short pulse of [^{3}H]-thymidine over one division cycle (grain count diminution method), the doubling time, defined by the authors as the time it takes cells to complete the S, G_2 and M phase (excluding the G_1 phase !), was determined for centroblasts, i.e., GC B cells in the dark zone of the germinal center, to range between $5-7$ hours. A stathmokinetic study (metaphase arrest method) presented by Zaitoun *et* al. [38] in 1980 concluded that the 'apparent' cell cycle time in the overall germinal center ranged from $15-17$ hours. In the same report two alternative methods, namely the fraction of labeled mitotic figures (FLM) technique and pulse labeling with tritiated thymidine, yielded somewhat different average estimates of 12.3 hours and 19.4 hours respectively. A similar study carried out by Zhang and co-workers in 1988 [36] reported for centroblasts a birth rate coefficient of 15 % per hour, which translates under common assumptions[1] into a generation time of $6-7$ hours (i.e., $100/15 = 6.6$). In 1991, Liu *et* al. [134] undertook an in-depth study of sites of B cell activation during primary and secondary immune responses in rats. The authors concluded from immunohistology examinations of *in vivo* BrdU pulse labeled spleen sections that B cells in the light zone were renewed by cells from the DZ every 7 hours and that the generation time of cells in the DZ was about 6 hours. Later, Hollowood and colleagues [60] carried out a thorough stathmokinetic study measuring GC cell birth rates at several stages during the primary and secondary immune response. They found that pre-existing GC in unimmunized animals had a mean cell birth rate coefficient of 3.3

[1] To interpret stathmokinetic data it is common to assume that the generation time equals the reciprocal of the mitotic entry rate coefficient. In principle this requires, in order to represent a valid approximation, a homogeneous and self-sustained proliferating cell population in steady state, in which in average half of the new-born cells die or leave the population immediately after birth (see Introduction).

Method	Data	Quantity	Estimate	Author	Ref.
grain count dimin.	1964	doubling time (DZ)	5-7	Hanna	[61]
FLM	1980	G1, S, G2, gen. time (GC)	6.0, 4.8, 1.4, 12.3	Zaitoun	[38]
stathmokinetic	id.	birth rate, gen. time (GC)	6% per h, 16.4	id.	id.
stathmokinetic	1988	birth rate, gen. time (DZ)	15% per h, 6-7	Zhang	[36]
BrdU and histo.	1991	gen. time (DZ)	6	Liu	[134]
stathmokinetic	1992	birth rate (established GC)	3.3 % per h	Hollowood	[60]
BrdU-DAPI	2007	gen. time (GC)	>12	Allen	[37]
BrdU-DAPI	2007	gen. time (GC)	6-8	Hauser	[24]

Table 3.1: Overview over cell cycle estimates from GCs in the literature. Values of the estimates are given in hours, if not otherwise indicated.

% per hour. Two days after primary immunization this coefficient decreased to 1.8 % per hour followed by an almost threefold increase up to 4.4 % per hour on day four. Thereafter, it returned to its initial level by day 7, i.e., 3.3 % per hour, and remained stable for the rest of the experiment. Notably no attempt was made in this article to estimate the generation time in GCs. However using common model assumption an apparent generation time in established GCs of 30–31 hours (i.e., $100/3.3 = 30.3$) can be derived. A more recent study by Allen *et* al. in 2007 [26] questioned previous low estimates on generation times of GC B cells. Based on DNA-BrdU pulse labeling, the authors pointed out that the average GC B cell-cycle time most probably lay above 12 hours in their experiments. Concurrently Hauser *et* al. [24] observed during a comparable study a considerable number of GC B cells reentering S phase after 6-8 hours, showing that at least some cells did complete part of the S, the G2, the M and the G1 phase during this time interval.

3.1.3 Theoretical insights from germinal center models

Since the seminal paper by Kepler and Perelson in 1992 [154], understanding affinity maturation has remained one of the central themes in the GC modeling field. By analyzing a deterministic differential equation model of a GC B cell population, divided into several affinity classes, Kepler at al. found, based on a numerical optimal control treatment, that a mutation schedule with brief bursts of high mutation rates interspersed between periods of mutation-free growth would lead, in their model, to most efficient affinity maturation. Relating these theoretical results to biology, it was suggested in this same paper that such a model could provide a framework within which the anatomy and kinetics of the germinal center reaction could be understood. They proposed that the derived optimal mutation schedule, coined cyclic re-entry, was realized in the GC either by cohorts of i) selected centrocytes reentering the DZ via the outer zone, ii) memory B cells re-entering the GC via the circulation, or iii) memory B cells colonizing a new germinal center. Later, this model was adapted to account for antigen decay [155], stochasticity [156], explicit centrocyte to centroblast conversion [157, 158] and mutation occurring at transcription or at replication [158]. A major insight from these additional modeling studies was that recycling remained in most scenarios a necessary feature to explain efficient affinity maturation.
A hint, that affinity maturation might however not be as efficient as previously thought, came from a study analyzing B cell receptor (BCR) sequence data from cells picked in single GCs during the anti-4-hydroxy-3-nitrophenyl-acetyl (NP) immune response. Radmacher and Kepler computed in 1998 [159] the mean waiting time τ_{key} for a key mutation[2] to appear in a GC of size 10000 as follows

$$\tau_{key} = (660 \times 10^{-3} \times 0.19)^{-1} \approx 8 \text{ hours}$$

[2] A recurring mutation in a certain immunoglobulin heavy and light chain rearrangement that is known to increase the affinity to NP several fold.

where 660 is the number of new cells produced per hour in established GCs [60], 10^{-3} is the mutation rate per base pair per cell division and 0.19 accounts for the fact that the key mutation involves a $G \to T$ transversion, which occurs at a rate of 19% per G mutation (see references in [159]). This estimate was in stark contrast to the estimate of 8.3 days obtained from a maximum-likelihood method based on their BCR sequence data. This finding was later termed the key-mutation discrepancy [145]. Several non-exclusive explanations have since been provided to explain why so few key mutations have actually been observed, despite the fact that cells carrying this mutation should frequently be generated in GCs. While Radmacher *et* al. argued that many cells bearing key mutations are overlooked due to stochastic selection, Kleinstein *et* al. [145] put forward the idea that blocking mutation are most important in preventing establishment of high-affinity B cells. Interestingly, if one adds to the above computation, as proposed in the original paper by Radmacher *et* al., a factor 0.145 which accounts for the key-mutation being in a cold-spot and a factor 0.25 which is an estimate for the probability that a lethal mutation is acquired concurrently with the key mutation, one computes $\tau_{key} \approx 9.1$ days, which is even longer than the estimate based on the sequence data. This makes clear how much the key-mutation discrepancy depends on quantities that are extremely difficult to estimate, and leaves open the intriguing question whether affinity maturation is highly efficient or inherently 'wasteful'.

Further preventing a clear answer to this question is the fact that any reasonable model of affinity maturation needs to define an affinity space. Specifying a realistic affinity space is however a very complex undertaking and remains an unsolved problem [154]. Usually GC models circumvent this by making some simplifying assumptions, e.g., a small number of affinity classes [154, 158], mutation decision trees specific for certain canonical hapten responses [145] and low-dimensional shape-space models [160]. Importantly any conclusions concerning the efficiency of affinity maturation in GCs remain dependent on these assumptions.

Another aspect of affinity maturation, which has received considerable attention in the field, is the selection process *per se*. Hypermutated GC B cells with increased affinity relative to their peers are usually assumed to receive more survival signal or are induced to expand more rapidly, and several mechanism have been devised to account for this preferential treatment. While some studies explored antigen masking [161] and cellular sorting [162], other groups investigated the role of FDC dynamics [155] and T-cell help [157]. In this context, fast take-over rates estimated from DNA sequence data collected during anti-hapten immune response became of interest, because these indicated high stringency in the selection process. More precisely, early GCs (day 10 post immunization) which were found to contain B cells bearing key-mutations, were constituted exclusively of key-mutants. This observation suggested that the latter had quickly outcompeted germline cells. However, as pointed out by Kleinstein *et* al., fast take-over rate estimates were based on data from a single GC (!) of unknown size, who had apparently been founded already by a key-mutant. Thus, instead of indicating fast take-over rates, the finding above could alternatively provide support for the hypothesis by Kepler and Perelson, in which selected GC B cells may emigrate from their follicle in order to colonize a new follicle. This would also be in line with the observation that GC seeding is an on-going process throughout the GCR [163].

Until recently the most important data for GC modelers, besides of DNA sequences, were the average numbers of GC B cells in spleen and lymph nodes during the GCR. These highly reproducible kinetics have extensively been used to validate population dynamics predicted by the models mentioned before [155, 156, 162]. In addition to growth, the termination of the GCR has been studied as well [164, 165], however conclusive results were difficult to obtain based on these kinetics alone. Besides of average GC B cell numbers in spleen and lymph nodes, GC volume distributions have also been subjected to model-based analysis, leading to the hypothesis that GCs might undergo sudden collapses due to e.g., the appearance of a key-mutant [163].

With the advent of two-photon microscopy, it became feasible to image GC B cells as they migrate and divide in their natural environment *in vivo* [24, 25, 37]. This prompted several groups to develop complex agent-based GC models able to reproduce simultaneously GC B cell migration pattern, cell

division, dark and light zone formation, affinity maturation, differentiation and T-B cell interactions [85, 160, 166]. These models were utilized to test the conclusions drawn from 2-photon microscopy studies, with the main insights that persistent random walk could explain well GC B cell migration and that imaging times and volumes in the two-photon microscopy experiments had been insufficient to estimate accurately transzonal migration rates.

3.1.4 Germinal center models of proliferation and migration

To conclude the introduction, we review the main GC models that have been proposed in the past, especially focusing on proliferation and migration, while leaving, for the sake of comparison, other important features, like e.g., affinity maturation and GCR initiation and termination, apart.

The so-called classical model of GC dynamics assumes that centroblasts divide rapidly in the DZ to give continuously rise to non-proliferating centrocytes, which move to the light zone, where selection is presumed to take place. Selected centrocytes then leave the GC to become memory or plasma cells. A schematic representation of this model is shown in Fig. 3.2 A, where the gray area represent the DZ, the white area is the LZ, straight arrows show cell migration and the closed circular arrows indicate cell proliferation[3]. This model is usually attributed to Ian MacLennan [152], however it also matches relatively closely the migration and proliferation pattern proposed in the study by Hanna (see Fig. 3.1 A and [61]).

Kepler and Perelson suggested, as discussed in the previous section, a modification to this simple model in which centrocytes are, besides of differentiating into plasma and memory cell, allowed to recycle back into the DZ (see Fig. 3.1 B and Fig. 3.2 B). This view was re-adopted almost fifteen years later by Allen *et* al. and others to interpret live-imaging data of GC B cells in murine lymph nodes (see Fig. 3.1 C and [37]). Several aspects however differed from the original hypothesis, namely in this updated model centrocytes recycled back directly through the DZ/LZ interface instead of taking an external route and proliferation took place both in the DZ and in the LZ (see e.g., Fig. 3.2 C)[4]. Migration rates were determined by Allen *et* al. as 4% per hour for cells migration from the DZ towards the LZ and 2% for cells migrating from the LZ towards the DZ. A similar behavior was observed by Hauser *et* al., which led these authors to conclude that migration, proliferation and also selection is predominantly intrazonal [24].

As mentioned before, Victora *et* al. followed migration of GC cells over longer time periods, which probably yielded more accurate cell flux estimates, compared to previous studies. They observed in their experiments that 15% of cells migrated per hour from the DZ to the LZ while in the same time only 1.5% (relative to the DZ) migrated from the LZ to the DZ (see Fig. 3.2 D). In addition, by determining the frequency of cells in G_1, S and G_2M phase in the DZ and in the LZ, they surprisingly found that cells in G_2M phase were virtually absent in the LZ. This indicates that cells which are in CC in the LZ, do not complete the CC therein (see e.g., Fig. 3.2 D). While it is currently unknown where these cells undergo mitosis and cytokinesis, one plausible assumption is that they return to the DZ. Together, this suggests that even though a sizable amount of cells in GCs progress through parts of their CC in the LZ, they always complete it in the DZ. Based on these and other findings, Meyer-Hermann *et* al. proposed that GC B cells selected in the LZ enter the cell cycle in the LZ, then migrate to the DZ, divide twice with a division time of 6 hours and then either leave the GC as plasma cells or return to the LZ for a further round of selection (see [85] and Fig. 3.2 D).

[3] Closed circular arrows have been used in previous GC studies to indicate intra-zonal migration [23,149]. Here they will be used exclusively to denote cell proliferation, meaning that dividing cells are completing a full cell cycle within the respective compartment. Open circles denote cells which complete only part of the cell cycle within the respective compartment.

[4] Biological models, in contrast to mathematical models, are typically vaguely defined. Fig. 3.2 C is an attempt to translate hypotheses expressed in [37] into a concrete mathematical model. It should not be regarded as 'the' model proposed in the paper, but rather one possible model, consistent with the statements made therein.

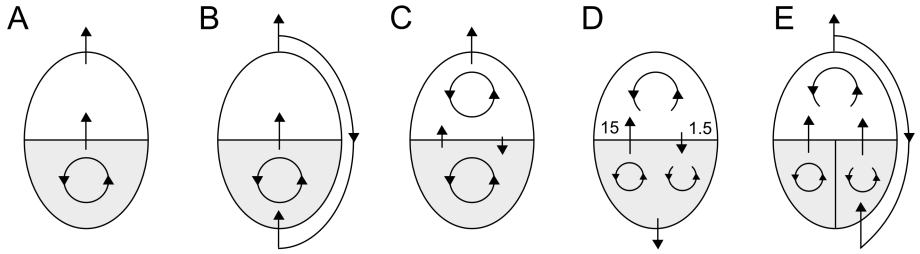

Figure 3.2: GC proliferation and migration models. A) The classical model of GC dynamics assumes that B cells in the DZ (gray area) proliferate (closed circle) rapidly, then leave the CC and migrate (straight arrow) to the LZ (white area) to get selected and become memory and plasma cells [152]. B) The cyclic re-entry model assumes that some of the B cells leaving the LZ can re-enter the DZ for further rounds of proliferation [154, 158]. C) An updated model in which centrocytes recycle back directly through the DZ/LZ interface and proliferation takes place both in the DZ and in the LZ [26]. D) A recent model based on data from photoactivation experiments presumes a large net flow of cells from the DZ to the LZ [85]. Notice that cells complete only part of their CC in the LZ (open circle) and that they are leaving the GC through the DZ. E) Possible extension of the model that will be analyzed in the last section of this chapter. Cells in the DZ belong either to a self-renewing (left) or a trafficking population (right). Only the DZ part of this model will be considered.

3.2 Results

Currently it is not clear, which of the models shown in Fig. 3.2 is most appropriate to describe proliferation and migration in GCs. While textbook sections on GC biology still favor, perhaps due to its simplicity, the classical model (see [34] and Fig. 3.2 A), the majority of investigators working in the field probably 'believe' in some degree of re-cycling inside GCs (see e.g., [24, 25, 37, 153, 154, 156–158] and Fig. 3.2 B-E). Despite this important paradigm shift in the community, the way how recycling is organized remains to be elucidated.

3.2.1 Birth rates, S phase labeling index and DZ:LZ ratios argue against a classical view on GC B cell proliferation

In a GC DZ cell population, as proposed by model A (see Fig, 3.2), with a birth rate coefficient of 15 % per hour, as measured by Zhang and co-workers for dividing cells in the DZ [36], and a 17 % fraction of cells in S phase, as measured by photoactivation [153] and [^{3}H]-thymidine incorporation [61], centroblasts would have to divide every 6.6 hours (i.e., under common assumptions, 100/15 = 6.6) and duplicate their entire genome in little more than one hour (i.e., 17% of 6.6 hours = 1.1 hours). This, i.e., S phase durations of 1.1 hours, is several times faster than previous estimates for S phase durations reported for rapidly proliferating murine cells *in vivo* and *in vitro* [10], including GC B cells [38].
Moreover, in order to accommodate under the classical model, both the birth rate coefficient measured in the DZ with the GC wide birth rate coefficient of 3.3 % per hour [60], a DZ:LZ ratio of approximately 1:5 (i.e., 3.3/15 = 0.22) is expected [159], a value ten times lower than the recently determined ratio of 2:1 [153] and still 5 times lower than the more conservative estimates of 1:1 deduced from immunohistology [167].
Both observations suggest that the 6-7 hour generation time estimate for centroblasts may have been based on a model too simple to account for the complex environment of the GCs. The analysis in the subsequent sections will on one hand further support this view and will on the other hand show

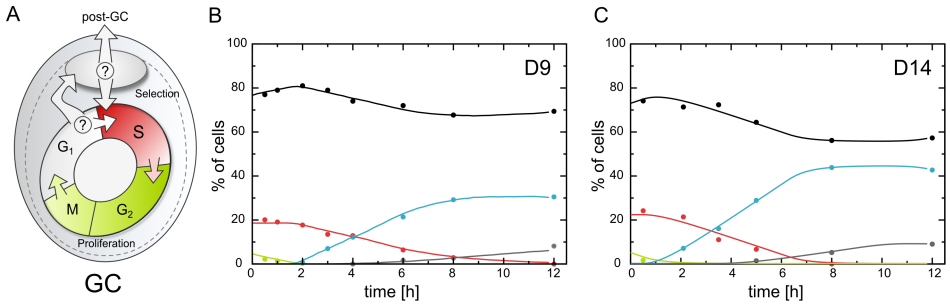

Figure 3.3: A proliferation and selection model of the GC. A) Pictorial representation of the proposed model. B cells in the GC either belong to the proliferating or selected phenotype. Two decision points, indicated by question marks (?), control their respective fate. At the end of the G_1 phase cells either enter a further round of DNA replication or undergo selection. When the selection process is completed, cells either differentiate back into the proliferating phenotype (re-entry) and start replicating their genome or disappear from the GC. B,C) Best fit of the model to kinetics extracted from BrdU pulse labeling data of GC B cells (B, [24]; C, [37]). Frequencies of labeled cells that have not divided since the pulse (red dots), of unlabeled cells initially in G_2M phase (green dots), of unlabeled cells initially in G_1 together with progeny of unlabeled cells born after the pulse (black dots), first generation progeny (counted from the pulse) of labeled cells (blue dots) and first generation progeny of labeled cells in S phase (gray dots) are approximated by the predictions of the model (lines).

that trafficking of a cycling GC B cell subpopulation through the DZ, in line with the cyclic re-entry model, provides a plausible explanation for both the high birth rates and the low percentage of cells in S phase in the DZ, without the need for extreme DNA replication rates.

3.2.2 Model-based analysis fails to detect a rapidly replicating subpopulation in GCs

A substantial number of cells in the GC dividing every 6-7 hours and capable of duplicating their whole genome in approximately one hour, should certainly leave a clear signature in experiments able to quantify cell cycle progression and DNA synthesis. Therefore we interrogate two independently published data sets from DNA-BrdU pulse-chase labeling experiments with an abstract GC B cell proliferation model in order to see if we can find traces of such a 'remarkable' cell population.
The model presumes (see Fig.3.3 A) that there are in the GC two major B cell phenotypes, one proliferating (in G1, S and G2M phase) and one undergoing selection (in G1 phase). Proliferating cells may 'exit' the cell cycle with a certain probability to undergo selection. After completing the selection process, cells either disappear from the GC through migration and death or differentiate back into the proliferating phenotype (re-entry). Where selection and proliferation take place, i.e., either in the DZ or in the LZ or in both, is not specified. The completion times for each phase, including selection, are assumed, as in Section 1.4.1, independent shifted exponentially distributed random variables depending on two parameters each (see Material and Methods for a rigorous mathematical definition and analytical predictions of the model).
The data sets we test for the presence of rapidly replicating cells in GCs consist of measurements of BrdU uptake and DNA content by flow cytometry (i.e., obtained with the DNA-BrdU pulse-chase labeling method) in a large number of single cells from murine lymph nodes, identified by several markers as GC B cells [24, 37]. Samples had been taken at multiple time points, ranging from 30 minutes up to 12 hours after administration of a single injection of BrdU. Mice had been immunized 9 days [24] and 14 days [37] before with (4-hydroxy-3-nitrophenyl)acetyl-chicken gamma globulin (NP).

We will denote these data sets in the following as the D9 and the D14 data set. More detailed experimental procedures are given in the respective publications.
As a preliminary validation, we minimize the residual sum of squares between the model and the data by adjusting the eight parameters that define the times cells spend in each phase and an additional parameter that regulates the probability of re-entry. The model assimilates well the empirical dynamics observed in D9 and D14 (see Fig.3.3 B-C).
To further interrogate these data sets, we rely on the Bayesian inference framework (see Section 3.3.4). As discussed before, the typical output from Bayesian inference, given data, a model and a properly chosen prior, is a so-called posterior distribution over the model parameters. The latter reflects an estimate of the uncertainty in the parameter values. The marginal posterior distributions for several model parameter summaries based on analytical average kinetics, derived in the Section 3.3.1, are represented in Fig. 3.4. The corresponding means of the marginal posteriors and 95%-credibility intervals are summarized in Table 3.2.
For the proliferating phenotype almost typical estimates for the duration of G1, S, G2M are obtained (see Table 3.2 and Table 3.1). The total cell cycle length of $10.7\,(7.9 : 13.5)$ hours and of $9.9\,(5.5 : 14.9)$ hours extracted from the D9 and the D14 data set are in good agreement within each other and match relatively closely the 9.3 hours measured in *in vitro* studies with activated proliferating B cells [42]. The cell cycle kinetics, especially the S and G2M phase, appear slightly accelerated in the D14 when compared to the D9 data set. The probability of re-entry, defined as the probability for GC B cells to differentiate back into a proliferating phenotype after selection is however somewhat lower in the D14 data set (mean: 6% (D9), 2% (D14); relative to cells that completed cytokinesis). The minimal cell cycle length is estimated as $7.99\,(4.32 : 10.94)$ hours and $6.4\,(2.63 : 10.01)$ hours for the D9 and D14 data set respectively. Notice however that there only very few if any cells which complete a full cycle that fast.
The phenotype under selection shows surprisingly extended selection periods which over-pass the cell cycle length by a factor of 3.0 (D9) and 1.5 (D14). Interestingly, cells in the D9 experiment, as indicated by Bayesian summary statistics, do take twice as long, $30.4\,(18.2 : 43.5)$ hours as compared to $15.4\,(0.0 : 27.2)$ hours for the D14 data set, to complete selection. The fraction of cells undergoing selection is estimated as $72\,(59 : 81)$ % (D9) and $52\,(13 : 80)$ % (D14).
To summarize, the estimates for the average generation time range for both data sets at around 10 hours, similar to what has been observed for activated B cells in vitro [42]. The GC wide S phase duration are estimated to last 7.5 hours and 6 hours for the D9 and D14 experiments. Importantly no bi-modal kinetics, e.g., a fast and a slow progressing S phase sub-population, can be detected. The selection process, as represented in our model, takes longer than the respective total cell cycle length, with a duration two times more extended in the case of the D9 data set when compared to the D14 data set. It should however be noted at this point that all the above estimates depend on our model assumptions, and as such should be interpreted with extreme care. Especially the inferences regarding selection, the total cell cycle length, and the probability to recycle are rather sensitive, as the data does not allow to distinguish cells which are in the selection phase from actively proliferating cells which are in G_1 of the cell cycle. Nevertheless, the estimates for the average S phase duration are relatively robust to model assumptions, which also provides the motivation for the title of this section.

3.2.3 Heterogeneous model explains both DZ cell cycle and migratory data without the need for short generation times

By analyzing GC population wide cell cycle kinetics, we can't find any indication for a substantial cell population in GCs dividing every $6-7$ hours, nor does a sizable fraction of cells complete the S phase within 1 hour. This however is expected under the classical model of GC B cell proliferation and given available cell cycle data. Therefore a biologically plausible alternative is required that can reconcile these observations instead.

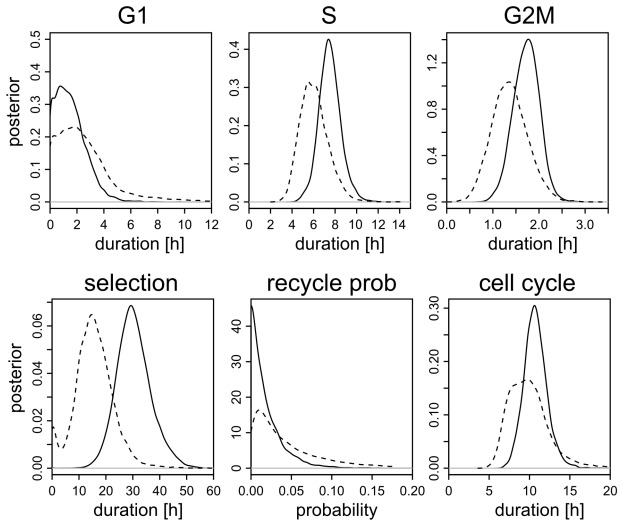

Figure 3.4: Bayesian Inference. Samples from the posterior density distribution over the model parameter are generated for D9 (solid line) and D14 (dashed line) using a Markov-Chain-Monte-Carlo (MCMC) approach (see Appendix, Algorithm). While the estimate for the average cell cycle duration ranges for both data sets around 10 hours, summary statistics suggest that selection lasts twice as long (30 hours (D9), 15 hours (D14)) in the D9 experiment.

	$\tau_{\mathrm{G_1}}$	τ_{S}	$\tau_{\mathrm{G_2M}}$	selection	cell cycle	$p_{re-entry}$
D9	1.5 (0:3.5)	7.5 (5.4:9.5)	1.7 (1.1:2.2)	30.4 (18.2:43.5)	10.7 (7.9:13.5)	0.06 (0:0.06)
D14	2.5 (0:6.9)	6.0 (3.7:8.6)	1.3 (0.6:2.1)	15.4 (0:27.2)	9.9 (5.5:14.9)	0.02 (0:0.32)

Table 3.2: Summary statistics (mean, 95%-credibility interval) of the computed marginal posterior density distributions. All values are given in hours, except the probability to differentiate back into the proliferating phenotype, which is unit-less.

Cell cycle dependent trafficking has repeatedly been proposed in the literature [37, 134, 153, 157], e.g., to explain GC related experimental findings like asymmetry in transzonal migration rates, the accumulation of cells in mitosis in the DZ and the increase of BrdU positive cells in the LZ after pulse labeling. In addition the latter mechanism was predicted by some 'stringent' forms of the cyclic re-entry model [24, 154], in which proliferating GC B cells, in order to undergo selection, traverse the LZ once per cell cycle. However the impact of GC trafficking on cell birth rates and derived generation time estimates, which is of special interest here, has received relatively little attention so far.

Experimentally, long-term ($\approx$ 6 hours) migration into and out of the DZ of single lymph node GCs has been assessed [153] and was complemented recently by an extensive modeling and simulation study [85]. The data reported in the primary study showed (our own interpretation, see Fig 3.5 B) that the percentage of activated cells in the DZ initially decreases by 40% in the first 4 hours following photoactivation, and then settles onto a plateau over the rest of the experiment. This indicates that about 40% of cells in the DZ of GCs are short-term visitors with the majority remaining in the DZ for less than 4 hours.

To understand the impact of such trafficking on stathmokinetic and related experiments in the DZ, and to see if we can reconcile the results from the previous section with birth rates and fractions of

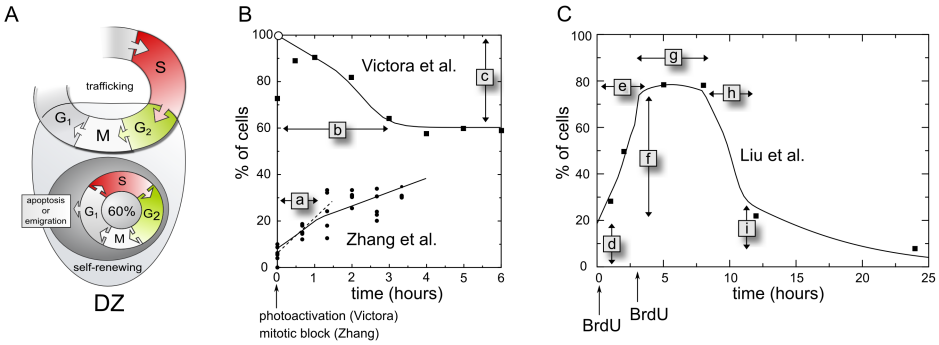

Figure 3.5: Heterogeneous cell cycle and trafficking model of the germinal center dark zone. (A) Besides a self renewing populations making 60% of cells in the DZ, a trafficking population that completes S phase outside, continuously enters the DZ in G2 phase, progresses to mitosis and undergoes cell division. After cell division the newly born cells, which are in G1 phase, remain in the DZ for some time before they emigrate. (B-C) Simultaneous fitting of the model (lines) to three published data sets: i) to the percentage of photo-labeled cells in the DZ activated in the DZ (B, black squares, adapted from Fig. S4.A and Fig. S4.B in [153]), ii) to the percentage of mitotic figures in the DZ after blocking cells in mitosis (B, black dots, reproduced from [36]) and iii) to the percentage of labeled cells in the DZ after BrdU pulse labeling (C, black squares, adapted from Fig. 5 in [134]). The direct read-outs are: **a**) the time until accumulation of mitotic cells prevents further immigration ($\approx$ 1.5 hours); **b**) the time it takes most trafficking photo-labeled cells to leave the DZ ($\approx$ 3 hours); **c**) the percentage of trafficking cells in the DZ ($\approx$ 40%); **d**) the percentage of S phase cells in the DZ ($\approx$ 17%); **e**) the time it takes until all trafficking cells are labeled, corresponding to b ($\approx$ 3 hours); **f**) the increase of BrdU labeled cells, of which 40% are due to the trafficking population ($\approx$ 60%); **g**) the time between the second pulse and the decrease in labeled cells, which equals the S phase duration of the trafficking population ($\approx$ 5 hours); **h**) the time it takes the labeled trafficking cells to leave the DZ, corresponding to b and e ($\approx$ 3 hours); **i**) the percentage of labeled cells that are lost due to label dilution over 12 hours ($\approx$ 20%). Also shown is the 15% slope based on linear regression of the first 3 time points (B, dashed line) proposed by Zhang *et* al.. To facilitate comparison the data sets by Liu *et* al. and Victora *et* al. were adapted and subsequently translated from their original units into %. The white dot at 100% in B represents our re-estimate of the no. of cells initially photo-labeled (see Materials and Methods, Data Adaptation).

cells in S phase reported in the literature, a cell cycle and trafficking model of the DZ, as depicted in Fig 3.5 A, is analyzed. For the DZ we assume a heterogeneous GC cell population[5] composed of cells with two different migratory and cell cycle characteristics, a self-renewing population making 60 % of all cells in the DZ and a second population that immigrates into the DZ after completing the S phase outside. Inside the DZ, cells appertaining to the second population, first enter mitosis, undergo cell division and then leave the DZ in G_1 phase. In the following, these cells, in S phase outside the DZ, and in G_2, M, and G_1 phase inside the DZ, will be referred to as the trafficking subpopulation. Finally half of the cells in the self-renewing population are assumed to emigrate or die, after completing the G_1 phase, as a consequence of homeostatic regulation. While the self-renewing population mimics the cell type that is expected under the classical model of GC dynamics, the trafficking population corresponds more closely to the phenotype predicted by the cyclic re-entry model [154, 158].
We confront this model simultaneously with three published data sets: i) the percentage of murine GC cells arrested in mitosis after administration of vincristine reported by Zhang *et* al. [36], ii) the percentage of photoactivated murine GC cells that remained in the DZ after being activated in the DZ presented by Victora *et* al. [153] and iii) the percentage of BrdU positive GC cells recorded over time after two BrdU pulses in rats by Liu *et* al. [134].
In order to capture the three types of experiments some additional assumptions have to be specified. Besides the well documented empirical facts that BrdU labels cells in S phase and that vincristine arrests cells in mitosis, BrdU labeled cells that divide for the first time are presumed to generate progeny with detectable amount of BrdU. In contrast later progeny are born unlabeled, e.g., due to label dilution. Similarly, photoactivated cells that divide during the 6 hours of the experiment also generate photoactivated progeny. Cell migration and hence trafficking are assumed to be incompatible with mitotic arrest. Therefore trafficking cells arrested in mitosis accumulate in the DZ after administration of vincristine, until the DZ starts growing in size. Then, due to e.g., a density-dependent G_2 checkpoint, further entry into mitosis is prevented. Finally, for the sake of simplicity, neither trafficking nor self-renewing cells that leave the DZ, return to the DZ during the course of the experiments.
For the analysis we proceed in the same way as with the proliferation and selection model in the previous section. After initial least-squares-fitting, whose result are shown in Fig. 3.5 B-C, Bayesian inference employing minimal prior knowledge is applied using for both methods analytical average kinetics (see Section 3.3.2). Due to two subpopulations, the relatively large number of parameters (16 in total, 2 populations with 8 parameters each, specifying the completion time in 4 phases) and limited information about the variables, it is not expected that all parameters can be uniquely identified. Instead, summary statistics capturing general traits of interest, namely the total cell cycle length of the self-renewing population, the transit time of the trafficking cell population and the average S phase duration for both phenotypes are extracted (see Fig. 3.6).
The total average cell cycle length for the self-renewing population is estimated to last between $11-18$ hours, close to the range of the $15-17$ hours estimated by stathmokinetic experiments in the overall GC [38]. The average S phase duration estimate lies around 5.3 hours for both populations. The latter estimates are somewhat lower than the overall S phase durations extracted from the D9 and the D14 data set, however again consistent with the 4.8 hours reported in [38]. Cells appertaining to the trafficking population are found to stay in the DZ for about 3 hours, a time during which they have, according to our model, to complete mitosis and undergo cytokinesis.
These results show that, given our model, neither rapid proliferation nor fast DNA replication is required to simultaneously explain empirical birth rates and percentages of cells in S phase in the DZ. In addition, besides of providing this proof of concept, our model, a hybrid between the classical and the cyclic re-entry model, explains naturally how the percentage of BrdU positive cells can raise by a factor of three in only 5 hours (!) in the experiments of Liu *et* al. (see Fig. 3.5 C) and suggests a new interpretation for the plateau-like kinetics observed after 2 hours and 4 hours in the studies by Zhang *et* al. and Victora *et* al. respectively (see Fig. 3.5 B). None of these observations can easily

[5] For simplicity T cells and macrophages are not considered in the model.

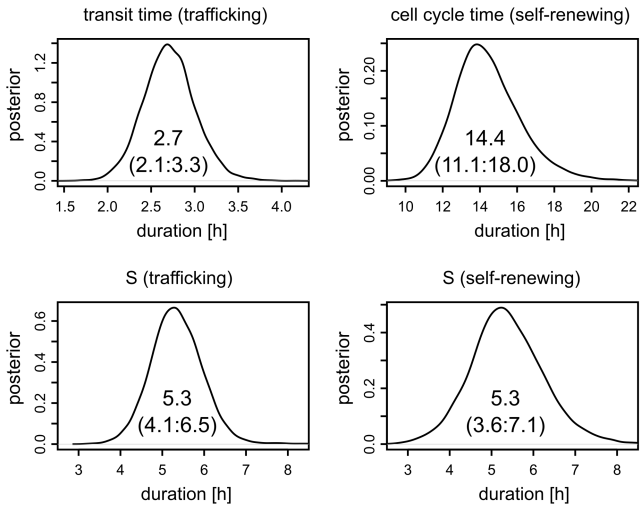

Figure 3.6: Marginal posterior densities given the model and the data as presented in Fig. 3.5. Shown are for the trafficking phenotype the posterior over the transit time, i.e., the time lapse between entry and exit, and the S phase duration. For the self-renewing phenotype the cell cycle length and the S phase duration are represented. For each density the average and the 95%-credibility interval are given under the respective graph.

be explained neither by the classical nor by a homogeneous cyclic-reentry model.
Intriguingly, the present analysis also indicates a substantial DZ immigration rate of 13% per hour (e.g., 40/3.0=13.3) of cells that complete S phase shortly before entry. In the light of the rising acceptance of the cyclic re-entry model in the field [85, 166, 168], this might not seem too surprising. However the fact that the emigration rates from the LZ to the DZ as measured by Victora *et* al. can only account for a small fraction (approximately 5%)[6] of this flux implies that many of the immigrating cells would have to enter the DZ, not as usually assumed, from the LZ but *via* a different pathway. We will elaborate on this point in more detail in the discussion.
As in the previous section, we emphasize on the fact that all conclusions presented herein depend on our model assumptions and the data, both of which are associated with a high degree of uncertainty. While our model is based on assumptions that have not been tested yet, the data we analyzed stems from different research groups, employing different protocols, and even different animal systems.

3.3 Materials and Methods

3.3.1 A GC proliferation and selection model

The GC proliferation and selection model depicted in Fig. 3.3 A is defined mathematically by a state transition model at the cellular level. The latter is completely specified by its states, transitions and rules that control the timing and the probabilities of the transitions. The set of possible states is given by four states inside the GC

$$s = \{\mathrm{G}_1,\ \mathrm{S},\ \mathrm{G}_2\mathrm{M},\ \mathrm{select}\},$$

and one auxiliary state 'exit' which is introduced to facilitate notations.

[6] From the 70-80 cells that were photoactivated in the LZ in the study by Victora *et* al., in average 5 migrated to the DZ in 4-6 hours. With a DZ:LZ ratio of 2, this corresponds approximately to 0.66% per hour (e.g., 0.5×5 cells /(75 cells $\times$ 5 hours) = 0.66) of immigrating cells into the DZ. This is 20 times less or 5% of the 13.3% per hour required to compensate for the emigration out of the DZ.

The transitions are of two types, a first type allows for a single fate only

$$\Lambda_1 = \{\mathrm{S} \xrightarrow{1} \mathrm{G_2M},\ \mathrm{G_2M} \xrightarrow{2} \mathrm{G_1}\},$$

while a second type allows for multiple fates

$$\Lambda_2 = \{\mathrm{G_1} \longrightarrow \begin{array}{c} \overset{1-R}{\nearrow} \\ \underset{R}{\searrow} \end{array} \begin{array}{c} \mathrm{S} \\ \mathrm{select} \end{array}, \mathrm{select} \longrightarrow \begin{array}{c} \overset{1-0.5/R}{\nearrow} \\ \underset{0.5/R}{\searrow} \end{array} \begin{array}{c} \mathrm{S} \\ \mathrm{exit} \end{array}\}.$$

The labels above the arrows indicate, for the events of the first kind, the factor by which a transition event is 'multiplied' as soon it occurs. For example, every cell that exits the $\mathrm{G_2M}$ phase corresponds to two cells entering the $\mathrm{G_1}$ phase, because cell division happens at this transition. For the events with multiple fates the labels represent probabilities instead. One can see, for instance, from the above expression that every cell that completes selection either leaves the GC with probability $0.5/R$ or initiates a new round of DNA replication with probability $1 - 0.5/R$.

In our model, cells which complete the $\mathrm{G_1}$ phase, enter either a further round of DNA replication or differentiate into the selected phenotype. The latter occurs with probability R, a parameter that, in order to avoid exponential growth of the proliferating cell population, can only take values between 0.5 and 1. The probability for re-entry $p_{re-entry}$ defined by $R - 0.5$ is such that, in average, half of the cells that complete the $\mathrm{G_1}$ phase enter the S phase at some later point. Again this is necessary to keep the GC population size constant over time.

In order to make predictions about the temporal evolution of this model, we have to specify the rules that control the time cells spend in each of the states. A simple, yet powerful approach, that we have already adopted in Chapter 1, is to assume state dwelling times that are shifted exponentially distributed. Recall that the latter depend on two parameters α and β, where α is the reciprocal of the rate of the exponential and β is the fixed delay. The mean dwelling time τ is $\alpha + \beta$ with standard deviation α and coefficient of variation $\frac{\alpha}{\alpha+\beta}$.

With this we have rigorously defined the GC proliferation and selection model that in total depends on 9 parameters, 8 of which specify the dwelling time in the four states inside the GC, and one determines the probability of re-entry.

A BrdU pulse-chase experiment then consists in labeling permanently, at a given time, all cells in the S state and follow their kinetics over time. The general strategy to obtain analytical predictions for DNA-BrdU pulse-chase labeling experiment based on this model consist, as seen for exponentially growing cell populations in Chapter 1, in setting up a steady state transition probability matrix, in solving the eigenvalue problem, and in computing the accumulated efflux of cell cohorts out of the phase in which they became initially labeled (e.g., the S phase for DNA-BrdU pulse-chase labeling experiments). From the latter, accumulated fluxes between subsequent phases can be derived, using the formalism of the inverse Laplace transform. Together the accumulated fluxes then allow to compute the kinetics of labeled cells, as they enter subsequent phases over time. Finally the kinetics of unlabeled cells can be derived from the solutions of labeled cells. We will in the following only sketch the major derivation steps that lead to the full solution, as a detailed derivation for exponentially growing cell populations was already provided in Chapter 1.

The steady state transition probability matrix which captures all the transitions in the model is given by

$$\mathcal{Q} = \begin{bmatrix} -\gamma_{\mathrm{G_1}} & 0 & 2\,\gamma_{\mathrm{G_2M}} & 0 \\ (1.0 - R)\gamma_{\mathrm{G_1}} & -\gamma_{\mathrm{S}} & 0 & (1 - 0.5/R)\gamma_{\mathrm{select}} \\ 0 & \gamma_{\mathrm{S}} & -\gamma_{\mathrm{G_2M}} & 0 \\ R\gamma_{\mathrm{G_1}} & 0 & 0 & -\gamma_{\mathrm{select}} \end{bmatrix} \tag{3.1}$$

where γ_i is $1/\bar{\tau}_i$ and R is the probability that a cell, at the end of $\mathrm{G_1}$ phase, differentiates into the

selected phenotype. The normalized eigenvector of $\mathcal{Q}$, which gives the steady state frequencies of cells in each phase is computed as

$$\begin{bmatrix} n_{\mathrm{G}_1} \\ n_{\mathrm{S}} \\ n_{\mathrm{G}_2\mathrm{M}} \\ n_{\mathrm{select}} \end{bmatrix} = \begin{bmatrix} 2\,\bar{\tau}_{\mathrm{G}_1} \\ \bar{\tau}_{\mathrm{S}} \\ \bar{\tau}_{\mathrm{G}_2\mathrm{M}} \\ 2\,R\,\bar{\tau}_{\mathrm{select}} \end{bmatrix} / T^* \tag{3.2}$$

where $T^* = (2\bar{\tau}_{\mathrm{G}_1} + \bar{\tau}_{\mathrm{S}} + \bar{\tau}_{\mathrm{G}_2\mathrm{M}} + 2R\bar{\tau}_{\mathrm{select}})$.
The fraction of BrdU positive cells that complete the initial S phase at time t after the pulse which occurred at t_0, can be derived from

$$\boldsymbol{\gamma}_{0\to1}(t) = \frac{\int_{-\infty}^{t_0} f_{\tau_0}(t-x)dx}{\int_{-\infty}^{t_0}(1-\int_0^{t_0-x} f_{\tau_0}(x^*)\,dx^*)\,dx}, \tag{3.3}$$

where 0 and 1 denote the phase S and G_2M respectively and f_{τ_0} is the shifted negative exponential density distribution describing the stochastic completion times specific to the S phase.
The total fraction of BrdU positive cells that completed the initial S phase since t_0 is then computed as follows

$$\Gamma_{0\to1}(t) = \int_0^t \boldsymbol{\gamma}_{0\to1}(x)\,dx = \begin{cases} t/\tau_0 & t < \beta_0 \\ 1 - \frac{\alpha_0 e^{\frac{\beta_0 - t}{\alpha_0}}}{\tau_0} & t \geq \beta_0 \end{cases}. \tag{3.4}$$

The Laplace transform of this expression is found as

$$\mathcal{L}_\omega(\Gamma_{0\to1}(t)) = \frac{1+\alpha_0\omega - e^{-\beta_0\omega}}{\tau_0\omega^2(1+\alpha_0\omega)}.$$

The fraction of BrdU positive cells that has completed G_2M phase and has thus divided at least once, can be derived using the properties of the inverse Laplace transform $\mathcal{L}^{-1}$ and the Laplace transform of the shifted exponential distribution which is $e^{-\beta\omega}/(1+\alpha\omega)$. We can write

$$\Gamma_{1\to2}(t) = \mathcal{L}_t^{-1}(\mathcal{L}_\omega(\Gamma_{0\to1}(t)) \times \frac{e^{-\beta_1\,\omega}}{1+\alpha_1\,\omega}), \tag{3.5}$$

which, with the help of modern algebra software gives

$$\Gamma_{1\to2}(t) = \begin{cases} \frac{\alpha_1 e^{\frac{\beta_1 - t}{\alpha_1}} - \tau_1}{\tau_0} & t < \bar{\beta} \\ 1 + \frac{\alpha_1 e^{\frac{\beta_1 - t}{\alpha_1}} - (\bar{\alpha}) e^{\frac{\bar{\beta} - t}{\alpha_0}}}{\tau_0} & \beta_0 > t \geq \bar{\beta} \end{cases}, \tag{3.6}$$

where $\bar{\alpha} = \alpha_0 + \alpha_1$, $\bar{\beta} = \beta_0 + \beta_1$ and the index 1 and 2 is short for G_2M and G_1 respectively.
Further accumulated fluxes can be derived using the following generic formula

$$\Gamma_{p\to p+1}(t) = \mathcal{L}_t^{-1}(\mathcal{L}_\omega(\Gamma_{0\to1}(t)) \times \prod_{k=1}^{p} \frac{e^{-\beta_k\omega}}{1+\alpha_k\omega}),$$

with the index k iterating over all intermediary phases. For example for $\Gamma_{3\to4}(t)$ which corresponds to the accumulated flux of labeled cells that divided once after the pulse and are entering G_2M, the index k iterates over the set $\{\mathrm{G}_2\mathrm{M}, \mathrm{G}_1, \mathrm{S}\}$.
We obtain the fate of labeled cells in various phases $n_i(t)$ mainly by straight-forward subtraction of

the accumulated fluxes, i.e.,

$$\begin{aligned}
n_0(t) =& n_0(1-\Gamma_{0\to1}),\\
n_1(t) =& n_0(\Gamma_{0\to1}-\Gamma_{1\to2}),\\
n_2(t) =& 2n_0(\Gamma_{1\to2}-\Gamma_{2\to3}),\\
n_3(t) =& 2n_0(1-R)(\Gamma_{2\to3}-\Gamma_{3\to4})+\\
&+2n_0(R-0.5)(\Gamma_{3\to6}-\Gamma_{6\to7}),\\
n_4(t) =& 2n_0(1-R)(\Gamma_{3\to4}-\Gamma_{4\to5})+\\
&+2n_0(R-0.5)(\Gamma_{6\to7}-\Gamma_{7\to8}),\\
n_6(t) =& 2n_0R(\Gamma_{2\to3}-\Gamma_{3\to6}).
\end{aligned} \tag{3.7}$$

Here the index set $\{0,1,2,3,4,5,6,7,8\}$ maps onto

$$\{\mathrm{S},\mathrm{G_2M},\mathrm{G_1^*},\mathrm{S^*},\mathrm{G_2M^*},\mathrm{G_1^{**}},\mathrm{select},\mathrm{S^{*s}},\mathrm{G_2M^{*s}}\},$$

where the star gives the number of divisions after the pulse, and the subscript s indicates that the flux corresponds to cell cohorts that passed through selection after the pulse. For example, the frequency of labeled cells in G1 that divided once after the BrdU pulse is given by the third row in Eq. 3.7, i.e., $n_{\mathrm{G_1^*}}(t) = 2n_{\mathrm{S}}(\Gamma_{\mathrm{G_2M}\to\mathrm{G_1^*}} - \Gamma_{\mathrm{G_1^*}\to\mathrm{S^*}})$.

Finally, in a real single pulse BrdU labeling experiment, not all of these populations can be readily identified. Frequencies that could be extracted from the data set we analyzed are labeled cells (f^{lu}), unlabeled cells initially in $\mathrm{G_2M}$ ($\mathrm{f}^{u}_{\mathrm{G2M}}$), unlabeled cells initially in $\mathrm{G_1}$ together with unlabeled cells that divided at least once (f^{u}_{G1}), first generation progeny of labeled cells in $\mathrm{G_1}$ phase (f^{ld}_{G1}) and first generation progeny of labeled cells that re-entered S phase ($\mathrm{f}^{ld}_{\mathrm{S}}$). Notice that the last two populations were considered as a single population in Chapter 1, because it is usually difficult to distinguish by FACS both populations in single BrdU-DNA pulse-chase labeling experiments. For the present chapter, we gated these subpopulations, despite the fact that our estimates are probably imprecise, because re-entry of labeled cells into S phase was an important observation in the original papers [24, 37].

All populations listed before can be constructed by adding several expressions defined in Eq.3.7 in the following way

$$\begin{aligned}
\mathrm{f}^{lu}(t) =& n_0(t)+n_1(t),\\
\mathrm{f}^{u}_{\mathrm{G2M}}(t) =& m_{\mathrm{G_2M}}(t),\\
\mathrm{f}^{u}_{\mathrm{G1}}(t) =& 1-(\mathrm{f}^{lu}(t)+\mathrm{f}^{u}_{\mathrm{G2M}}(t)+\mathrm{f}^{ld}_{\mathrm{G1}}(t)+\mathrm{f}^{ld}_{\mathrm{S}}(t)).\\
\mathrm{f}^{ld}_{\mathrm{G1}}(t) =& n_2(t),\\
\mathrm{f}^{ld}_{\mathrm{S}}(t) =& n_3(t)+n_4(t)+n_6(),
\end{aligned} \tag{3.8}$$

Here the right hand side in the second line corresponds to

$$m_{\mathrm{G_2M}}(t) = n_{\mathrm{G_2M}}\times\begin{cases}1-t/\tau_{\mathrm{G_2M}} & t<\beta_{\mathrm{G_2M}}\\ \alpha_{\mathrm{G_2M}}\exp[(\beta_{\mathrm{G_2M}}-t)/\alpha_{\mathrm{G_2M}}]/\tau_{\mathrm{G_2M}} & t\geq\beta_{\mathrm{G_2M}}\end{cases}, \tag{3.9}$$

which equals the first line in Eq. 3.7, however for the case that we would have labeled cells in $\mathrm{G_2M}$ and not in S phase.

These are the equations that we use for both least-squares fitting of the model with data and Bayesian inference. Because in the experiments apparently no cells complete a second S phase in 12 hours (the duration of the experiments), $\Gamma_{6\to7}$ and $\Gamma_{7\to8}$ in Eq. 3.7 are set to zero in order to facilitate the analysis.

3.3.2 A DZ proliferation and migration model

The DZ model shown in Fig. 3.5 A is defined formally, as in the case of the GC model, by a state transition model. Ignoring specific experimental conditions, e.g., labeling, mitotic block or photoactivation, we consider four phase-related states $s^{tr} = \{\mathrm{S}^{tr}, \mathrm{G}_2^{tr}, \mathrm{M}^{tr}, \mathrm{G}_1^{tr}\}$ and $s^{sr} = \{\mathrm{S}^{sr}, \mathrm{G}_2^{sr}, \mathrm{M}^{sr}, \mathrm{G}_1^{sr}\}$, for the trafficking ($tr$) and the self-renewing (sr) population respectively.
The transitions for the trafficking phenotype Λ_1^{tr} are given by

$$\{\mathrm{S}^{\mathrm{tr}} \xrightarrow{1} \mathrm{G}_2^{\mathrm{tr}},\ \mathrm{G}_2^{\mathrm{tr}} \xrightarrow{1} \mathrm{M}^{\mathrm{tr}},\ \mathrm{M}^{\mathrm{tr}} \xrightarrow{2} \mathrm{G}_1^{\mathrm{tr}},\ \mathrm{G}_1^{\mathrm{tr}} \xrightarrow{1} \mathrm{exit}\},$$

while for the self-renewing phenotype we have

$$\Lambda_1^{sr} = \{\mathrm{S}^{\mathrm{sr}} \xrightarrow{1} \mathrm{G}_2^{\mathrm{sr}},\ \mathrm{G}_2^{\mathrm{sr}} \xrightarrow{1} \mathrm{M}^{\mathrm{sr}},\ \mathrm{M}^{\mathrm{sr}} \xrightarrow{2} \mathrm{G}_1^{\mathrm{sr}}\}$$

as well as

$$\Lambda_2^{sr} = \{\mathrm{G}_1^{\mathrm{sr}} \longrightarrow \begin{array}{l} \overset{0.5}{\nearrow}\ \mathrm{S}^{\mathrm{sr}} \\ \underset{0.5}{\searrow}\ \mathrm{exit} \end{array}\},$$

where exit denotes again the auxiliary state introduced to simplify the notation. Notice that cells in S^{tr} are, according to our model, located outside the DZ. These cells only enter the DZ, after completing DNA synthesis.
The dwelling time in each state are presumed as before a random variable defined by a shifted negative exponentially distributed probability density function. Consequently the steady state n_i, the fluxes $\Gamma^i_{p\to p+1}(t)$ and kinetics $n_p^i(t)$ can be derived as previously outlined. The additional index i in these expressions denotes the phase in which cell cohorts initially become labeled. Specifying this was not required when we described DNA-BrdU pulse-chase labeling experiments because cells always became positive in the same phase. Here, as we shall see below, cells in all phases may become labeled or contribute to observable kinetics. Together these expressions are used to concisely translate domain knowledge and assumptions about experimental conditions into mathematical models of stathmokinetic, photoactivation and BrdU pulse labeling experiments.
We will begin with describing the photoactivation experiment in the case of the transmigratory population. Photoactivation is independent of cell cycle position, therefore cells in all phases become activated. The kinetics of photoactivated cells in the DZ depend on the time cells remain in the DZ, which is given for trafficking photoactivated cells (n^{tr}_{photo}) by the following expression

$$\begin{aligned} n^{tr}_{photo}(t) = {} & n_{G_2}(n_0^{G_2}(t) + n_1^{G_2}(t) + 2n_2^{G_2}(t)) + \\ & + n_M(n_0^M(t) + 2n_1^M(t)) + \\ & + n_{G_1}(n_0^{G1}(t)), \end{aligned} \tag{3.10}$$

where e.g., the first row corresponds to cells that become activated in G_2 phase. These cells have to go through mitosis, divide and enter G_1 phase before they can leave the DZ. Notice that the third term in the first row and the second term in the second row are multiplied by a factor two, which accounts for the fact that cells divide, after completing mitosis.
Stathmokinetic studies follow the percentage of mitotic figures after blocking cells in mitosis. Because in our model we presume that cells blocked in mitosis are unable to migrate, after administration of vincristine only trafficking cells in G_1 will continue to leave the DZ. Due to the constant immigration rate from outside these cells are replaced by cells blocked in mitosis as soon as the experiment starts. As long as there are G_1 cells leaving the DZ, it is reasonable to assume that the volume occupied by all the cells in the DZ remains constant (i.e., two small cells in G_1 are replaced by on 'big' cell blocked in mitosis). When most G_1 cells left the DZ, which should be close to τ_{G_1} after injection of

vincristine, then the volume increases and the DZ will appear more crowded. Cells in our model are assumed to react to this change in volume by preventing immigration of further cells. For the fraction of transmigratory cells in mitosis we obtain for the time before τ_{G_1}

$$n^{tr}_{mitosis}(t) = \frac{n_{\mathrm{M}} + \frac{n_{\mathrm{G}_1}}{2}\,\Gamma^{\mathrm{G}_1}_{0\to 1}(t)}{n_{\mathrm{G}_2} + n_{\mathrm{M}} + n_{\mathrm{G}_1} - \frac{n_{\mathrm{G}_1}}{2}\,\Gamma^{\mathrm{G}_1}_{0\to 1}(t)} \tag{3.11}$$

where the first term in the nominator equals the initial fraction of cells in mitosis and the second term describes the replacement of the G_1 population. For the time after τ_{G_1}, we have instead

$$n^{tr}_{mitosis}(t) = \frac{n_{\mathrm{M}} + \frac{n_{\mathrm{G}_1}}{2}\,\Gamma^{\mathrm{G}_1}_{0\to 1}(\tau_{G_1}) + n_{\mathrm{G}_2}\,\Gamma^{\mathrm{G}_2}_{0\to 1}(t-\tau_{G_1})}{n_{\mathrm{G}_2} + n_{\mathrm{M}} + n_{\mathrm{G}_1} - \frac{n_{\mathrm{G}_1}}{2}\,\Gamma^{\mathrm{G}_1}_{0\to 1}(\tau_{G_1})}. \tag{3.12}$$

Notice that after τ_{G_1} only cells in G_2 enter mitosis, while the total number of trafficking cells remains constant.

For the fraction of BrdU positive cells after a first pulse at time $t_1 = 0$ we obtain for the trafficking cells

$$n^{tr}_{BrdU\,1}(t) = \frac{n_S(n^S_1(t) + n^S_2(t) + n^S_3(t))}{n_{\mathrm{G}_2} + n_{\mathrm{M}} + n_{\mathrm{G}_1}}. \tag{3.13}$$

The same applies in principle for the second pulse after time $t_2 > t_1$ (which is 3 hours later in the experiments by Liu *et* al.)

$$n^{tr}_{BrdU\,2}(t) = \frac{n_S(n^S_1(t) + n^S_2(t) + n^S_3(t))}{n_{\mathrm{G}_2} + n_{\mathrm{M}} + n_{\mathrm{G}_1}}. \tag{3.14}$$

However due to the fact that some cells become double-labeled, we can not simply take the sum of these two expression to compute the total fraction of BrdU positive cells. We need to subtract from this sum the double-labeled cell cohort, which does follow, due to different initial conditions, different kinetics when compared to the single labeled cell cohorts.

Solving the accumulated flux of double labeled cells leaving the initial phase we find

$$\Gamma_{0\to 1}(t) = \begin{cases} t/\tau_0 & t + t_2 < \beta_0, \\ & \beta_0 > t_2 \\ 1 - \frac{\alpha_0 e^{\frac{\beta_0 - t_2 - t}{\alpha_0}}}{\tau_0} & t + t_2 \geq \beta_0, \\ & \beta_0 > t_2 \\ \dfrac{\alpha_0 e^{\frac{-\beta_0 + t_2 + t}{\alpha_0}}(e^{\frac{t}{\alpha_0}} - 1)}{\tau_0} & \beta_0 \leq t_2 \end{cases} \tag{3.15}$$

which corresponds to the flux of cells that are labeled by the first pulse at t_1 and remain in S phase until t_2. The Laplace transform of this expression is computed as

$$\mathcal{L}_\omega(\Gamma_{0\to 1}(t)) = \begin{cases} \frac{1 + \alpha_0\omega - e^{-(\beta_0 - t_2)\omega}}{\tau_0\omega^2(1+\alpha_0\omega)} & \beta_0 > t_2 \\ & , \\ \frac{\alpha_0 e^{(\beta_0 - t_2)/\alpha_0}}{\tau_0\omega(1+\alpha_0\omega)} & \beta_0 \leq t_2 \end{cases} \tag{3.16}$$

from which all the fluxes and kinetics can be derived as before. The fraction of double labeled cells is

then given by

$$n^{tr}_{BrdU\,(1+2)}(t) = \frac{n_S(m^S_1(t) + m^S_2(t) + m^S_3(t))}{n_{\mathrm{G}_2} + n_{\mathrm{M}} + n_{\mathrm{G}_1}}, \tag{3.17}$$

where we used $m(t)$ for the new kinetics to distinguish them from the previous solutions.
Finally we get for the time t after the second pulse the fraction of BrdU positive cells by summing cells labeled in the first and second pulse and subtracting the double labeled cell as follows

$$n^{tr}_{BrdU}(t) = n^{tr}_{BrdU\,1}(t^*) + n^{tr}_{BrdU\,2}(t) - n^{tr}_{BrdU\,(1+2)}(t^*), \tag{3.18}$$

where $t^* = t + t_2$, which accounts for the circumstance that cell cohorts get labeled at different times. The kinetics for the self-renewing population are derived using similar logic. Because we assume that photoactivated cells generate photoactivated progeny, the number of photoactivated cells in the self-renewing population is expected to remain constant during the course of the experiment, i.e.,

$$n^{sr}_{photo}(t) = const.$$

The fraction of cell arrested in mitosis $n^{sr}_{mitosis}(t)$ is found as follows

$$\frac{n_M + 0.5 \times n_{G_1}\Gamma^{G_1}_{2\to 3}(t) + n_S\Gamma^{S}_{1\to 2}(t) + n_{G_2}\Gamma^{G2}_{0\to 1}(t)}{1 - 0.5 \times n_{G_1}\Gamma^{G_1}_{2\to 3}}. \tag{3.19}$$

Here the different fluxes in the nominator describe how the respective cell cohorts enter mitosis. The factor 0.5 accounts for homeostatic regulation by which half of the cells that pass through G_1 emigrate or die in order to maintain the population size constant.
The expression for the fraction of BrdU positive cells in the self-renewing population are derived as outlined before. Two simplification are however applied in order to facilitate the derivation. First the G_2 and M phase are pooled together, as in the germinal center selection and proliferation model discussed earlier with parameters $\beta_{G_2M} = \beta_{G_2} + \beta_M$ and $\alpha_{G_2M} = (\alpha^{-1}_{G_2} + \alpha^{-1}_{M})^{-1}$. Second we neglect initially labeled cells that during the time in between the two pulses (i.e., 3 hours) reentered S phases after having completed the G_2, M and G_1 phase. The latter assumption ensures that double labeled cells are uniquely constituted of cells that became labeled by the first pulse at t_1 and had remained in S phase until t_2.
Finally, in order to match experimental conditions, the trafficking and self-renewing phenotypes are pooled into a single DZ cell population. This is straightforward as the kinetics can be added by a weighted sum in which the weights correspond to the respective fraction of cells in each sub-population. Care however has to be taken in the stathmokinetic experiments, because the fractions are changing over time.

3.3.3 Data Adaptation

In order to facilitate the simultaneous comparison of the analyzed data sets with the predictions from the DZ model, the kinetics measured by Liu *et* al. and Victora *et* al. are translated from their original units, number of cells per area and number of cells respectively, into percentage. This translation, seemingly straightforward, requires unfortunately some non-trivial transformations, which we will describe in the following.
The number of BrdU positive cells counted by Liu *et* al. after BrdU pulse labeling (Fig. 5 in [134]) were originally reported in number of cells per square millimeter. Because neither the density of cells in GCs nor scale bars on microscopy images were provided in the article, the scaling factor, which is necessary to transform by division the original units into %, cannot be determined. As a work-around we choose the cell density (2100 cells/mm^2) such that the reported range of measurements stays just

below the 100% line. This results in average in approximately 80% of labeled cells between 6 and 12 hours after administrating the first BrdU pulse, approaching as close as possible the observation by the authors 'that the centroblasts were all labeled within 6 h of injecting BrdU' [134]. It also leads to a best-fit of the model to the data with a total percentage of cells initially in S phase of 18% (see Fig. 3.5 C, at time zero). This is close to the 17% that had previously been estimated by Hanna *et* al. using [^{3}H]-thymidine and more recently by Victora *et* al. employing direct photoactivation of DZ cells and subsequent cell cycle analysis. Increasing the cell density (2400 cells/mm^2) such that 70% instead of 80% of cells are labeled between 6 and 12 hours after BrdU administration, does not change the estimates for the parameters of trafficking subpopulation but increases and decreases by circa one hour the estimates (marginal posterior mean) for the cell cycle length and for the S phase duration of the self-renewing population respectively.

For the kinetics of photoactivated cells in GCs measured by Victora *et* al. a different issue prevents the direct translation from number of cells into %. Due to technical reasons, photoactivated cells in these experiments suffered from a substantial spatial overlap immediately after activation, preventing accurate counting of activated cells at this point. According to the authors this lead to a systematic underestimation of the number of cells that became initially activated. The observed overlap apparently equilibrated after half to one hours after photoactivation, when activated cells had migrated sufficiently in random directions, such that single cells could unequivocally be identified.

Due to the apparent bias in the original value we considered to discard the measurement from the first time point and derive a new estimate through extrapolation using simple linear regression (see Fig. 3.7). This yields an estimate for the number of cells initially activated close to 51 cells instead of the 37 cells reported. While arguably suboptimal, we feel for several reasons that this approach yields a more plausible estimate than the value reported.

First, the original value would suggest that, in order to explain the increase in the average number of photoactivated cell by a factor of 1.4 in one hour (e.g., 53 cells/37 cells = 1.4), at least 40% of the activated cells would have to divide within the same time interval (e.g., $0.6 + 2 \times 0.4 = 1.4$). While activation induced proliferation may be possible, it is unlikely that cells complete S phase and mitosis so rapidly.

Secondly, under homeostatic conditions, it can be shown on theoretical grounds that the mean number of photoactivated cells in the DZ should only decrease or remain constant in time but never increase. The original value suggests however an average increase of photoactivated cells in the DZ of ca 24% in one hour (e.g., $46/37 = 1.24$)[7] while our new estimate indicates a decrease of 10% (e.g., $46/51 = 0.90$).

The reasoning why the number of photoactivated cells should only decrease or remain constant in time is as follows. The kinetics of the number of cells in the DZ, denoted here by n, can be described, similar to the model developed in [153], by a very general ODE as follows

$$\dot{n} = an - dn + bm$$

where an is the division rate, dn is cell loss due to death and emigration, bm is the immigration rate which depends on a positive rate coefficient b and the number of cells outside the DZ that 'await' immigration into the DZ, denoted here by m. Under homeostatic condition $\dot{n} = 0$. Given that photoactivation in the DZ does not significantly alter neither proliferation nor migration nor apoptosis, this ODE equally holds for photoactivated cells n_{photo}. From this follows that $\dot{n}_{photo} \leq 0$ immediately after photoactivation because $m_{photo} \approx 0$.

The number of photoactivated cells in the DZ which had been activated in the DZ is finally derived by subtracting the number of photoactivated cells counted in the LZ (Fig S4.B in [153]) from the total number of photoactivated cells (Fig S4.A in [153]). Percentages relative to the cells that became initially activated are obtained by division with the estimated initial number of cells, i.e., 51, and multiplication by 100.

[7] 46 was the number of activated cells counted in the DZ one hour after activation.

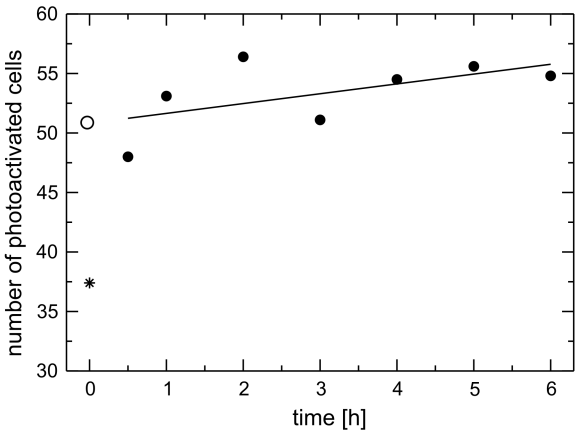

Figure 3.7: Data Adaptation. Number of photoactivated cells activated in the DZ (reproduced from Fig. S4.A in [153]). At time = 0 substantial overlap between photoactivated cells lead to underestimation of the experimental value (star, [153]). We extrapolated the number of cells photoactivated at time = 0 (open circle) using linear regression based on the remaining data points (line).

3.3.4 Bayesian Inference

Bayesian inference is carried out almost as outlined in Chapter 1 (see Section 1.5.2). As before, improper priors uniformly distributed over the positive real number are assumed for most parameters in both models. However for the parameter R in the GC model, a uniform density in the interval $[0.5, 1]$ is assigned in order to assure homeostatic control, and prior knowledge is introduced into the analysis of the DZ data by using a prior over the G2M duration for the self-renewing population (gaussian distributed with $\mu = 1.7$ and $\sigma = 0.28$, truncated at values smaller than 10^{-10}). The latter is the posterior estimated for the average duration of the G2 and M phase based on the BrdU pulse labeling data (D9). This serves to stabilize the MCMC chain and to maintain the parameters in physiological ranges despite the limited amount of data. It corresponds to the fair assumption that the self-renewing subpopulation progresses through the G2M phase at a velocity similar to an average cell in the GC.
The likelihood function as well the method to ensure convergence of the chains are identical to those used in Chapter 1 (Section 1.5.2).

3.4 Discussion

The classical model of the GC states that B cells in the DZ proliferate rapidly before they exit the cell cycle and migrate to the LZ. However by interrogating two data sets from DNA-BrdU pulse-chase labeling experiments with a simple proliferation and selection model we couldn't find any indication for an especially fast dividing subpopulation in the GC. This led us to hypothesize that the high birth rates measured in the DZ, which had partially fueled the widely held notion of rapidly dividing cells at this site, were not due to short generation times but were the consequence of cycling GC B cells through the DZ. To test this hypothesis, we compared a heterogeneous cell cycle and trafficking model of the DZ with data from three studies which had measured migration and cell cycle kinetics in the DZ using very distinct experimental techniques. We found that this hypothesis, in contrast to rapid proliferation, was consistent with these data sets. Although, this does not prove, especially in the light

of the high uncertainty associated with the analyzed data sets, the existence of such a population in the DZ, it provides an interesting explanation for both the high birth rates and the low frequency of cells in S phase measured in the DZ.

While our DZ model does not specify the origin of the DZ immigrants, it requires the incoming cells to be in cell cycle. Notably the number of cells migrating from the LZ into the DZ as measured by Victora *et* al. can only account for a small fraction of the expected flux (see Section 3.2.3). Therefore a cell population in cell cycle that continuously enters the DZ not from the LZ but *via* a different route appears to be necessary to explain these observations. Interestingly Kepler and Perelson [154], in their seminal paper, hypothesized that GC B cells either re-enter the DZ through the outer zone, or return as memory cells from the circulation. Specific B cells in cell cycle that could potentially join lymph node germinal centers have been observed in large numbers in the sub-capsular sinus in a recent study by Kerfoot *et* al. [169]. Furthermore Phan *et* al. [150] tracked B cells interacting with SCS macrophages which subsequently carried antigen into the GCs. And Schickert *et* al. [25] showed that GC are open structures that can be visited by both follicular and high-affinity antigen-specific B cells. A definitive answer to the question whether cycling B cells from the SCS continuously traffic through GCs could in principle be obtained experimentally relying on techniques similar to those developed by Victora *et* al.. Yet, what could be the role of such trafficking? Among a number of possible scenarios, trafficking back and fourth to the adjacent SCSs could favor replenishment of specific antigen consumed by GC B cells during the selection process.

Several other aspects of the DZ model deserve further discussion. In our model we assume, for the sake of simplicity, that cells enter the DZ just before of just after initiating the G_2 phase of the cell cycle. This coincides with cell cycle specific activation of B cell receptor signaling of GC B cells, as was shown in a recent report by Khalil *et* al. [170], suggesting, given that our model were appropriate, that immigration or subsequent progression through cytokinesis could be affinity dependent. Another important feature of the model is the self-renewing population. Although highly speculative, the latter could play the role of a bi-potent 'stem-cell-like' population inside GCs. Distinguishing characteristics for example of hematopoietic stem cells (HSC), the precursors of B and T lymphocytes, are niche-dependent self-renewal and production of more differentiated progeny. This fits well with DZ restricted cell division (i.e., cytokinesis) and the still poorly understood differentiation into plasma and memory cells associated with GC B cells. Perhaps differentiation (i.e., into plasma and memory cells) represents, as it has been shown for HSC, the default pathway for GC B cells and self-renewal is a property that is reserved to a privileged subpopulation in the DZ. Interestingly both HCS and GC B cells respond to CXCL12, a chemokine that is produced by mesenchymal stem cells in the bone marrow [171] and by stromal cells in the DZ of GCs [172]. Finally it should be noted that only 40% of cells in the DZ model belong to the self-renewing population. This could in principle explain, at least for the DZ, both the predominantly intrazonal mode of migration that has been observed during live-imaging experiments in murine lymph node GCs [24] and the 28% DZ cells moving towards the LZ along relatively straight paths [168].

A mathematical model describing cell cycle progression and migration from the dark to the light zone and back into the dark zone including a quiescent cell fraction had been developed several years ago to interpret part of the BrdU-DAPI pulse labeling data analyzed herein (see S5 in [24]). The model was represented by a system of differential equations, assuming for the DZ three successive stages, the S phase, the G_2M phase, and finally after division, a 'DZ' population. From this 'DZ' population, cells were assumed to migrate towards the 'LZ' population, where they were either leaving the GC through death or emigration, or returned to the S phase in the DZ. Progression in cell cycle and migration was described by constant rates (i.e., exponentially distributed waiting times), and the percentage of cells in S phase in steady state was fitted to experimental data. This then yielded estimates for the migration rates from the DZ towards the LZ, and from the LZ towards the DZ. The major differences between the present modeling approach and the one adopted in this former study is that the full kinetics of BrdU pulse labeled cells and not the steady state fraction of BrdU positive

cells (i.e., measured immediately after the pulse) were used in order to interrogate GC B cells. Despite the differences, these models share many similarities, and together highlight the potential of including heterogeneity and especially explicit cell cycle phases into models of GC dynamics.
Another more recent simulation model of the GC, proposed by Meyer-Hermann *et* al. [85], and largely based on data from the photoactivation experiments described before, presumes for the DZ, a homogeneous population of cells which, after immigration from the LZ, divide asymmetrically twice with an average generation time of 6 hours. From the progeny 75% differentiate into plasma and memory cell and leave the DZ, while 25% return to the LZ. It would be informative to see under which conditions this model would reproduce the kinetics shown in Fig.3.5 B-C.
In conclusion, our analysis suggests that the widely held notion of rapidly proliferating GC B cells in the DZ is incompatible with available data, and that a cycling GC B cell subpopulation undergoing cytokinesis while trafficking through the DZ could provide an alternative explanation for the high birth rates measured at this site. Even though faster cell cycle progression could potentially speed-up affinity maturation, and thus represent a fitness advantage, a trade-off may exist between quantity and quality, namely between cell production and faithful DNA replication or efficient affinity selection.

4 General discussion

In this last chapter, the main findings presented in this thesis are recapitulated. Then follows a discussion, which critically reviews the assumptions and limitations underlying our model-based inference approach. Finally, future directions and hypothetical applications are considered.

4.1 Summary

One of the aims of this thesis was to develop a mathematical framework, which would permit to characterize and infer phase completion times from DNA nucleoside analog pulse-chase labeling experiments. We started this 'journey', in **Chapter 1**, by setting up a stochastic phase-resolved cell cycle model, in which the phase durations for the G_1, the S and the G_2/M phase were assumed to be independent random variables, distributed each according to a shifted exponential density distribution (see Fig. 1.1). Analytical analysis was carried out for this model, steady state frequencies of cells in each phase were derived, and a numerical method which efficiently computes the asymptotic growth rate was implemented. Predictions for the division time distributions were fitted to two published *in vitro* data sets of dividing cell populations, which showed that the model could reproduce with reasonable accuracy these empirical measurements of cell cycle progression *in vitro* (see Fig. 1.2). With this model, we then analyzed and interpreted data from DNA-BrdU pulse-chase labeling experiments. By comparing the theoretical predictions with the pulse-chase data from a glioma cell line and a Chinese hamster cell line, we found that the model could also approximate well these more complex kinetic data sets (see Fig. 1.4). A subsequent Bayesian analysis revealed however, that some model parameters, especially those related to the G_1 phase, could not be fully identified based on the available data sets (see Fig. 1.6).
We addressed this problem in **Chapter 2**, by first identifying the source of the uncertainty in the parameter values. We realized that measurement noise was only partially responsible for the poor resolution, and showed *in silico* that the positions of the support points may have, in this system, a major impact on the quality of the estimates (see Fig. 2.2). To find the positions of the support points which would minimize the uncertainty in the parameter estimates, we relied on the theory of D-optimal design. This revealed that two experiments, with tree sampling points each, were sufficient to identify test parameter sets in *in silico* experiments in which either measurement noise was very low or the number of replicates was very high (see Fig. 2.5). However, under more realistic noisy conditions, the number of replicates required to infer the parameters at a reasonable precision, turned out to be prohibitively high (see Fig. 2.6). This led us to test *in silico* a modification of the prevailing DNA nucleoside analog pulse-chase labeling protocol, in which a second pulse with a different nucleoside analog is given shortly before cells are collected for fixation (see Fig. 2.7). To quantify the gain in information from this second pulse, artificial data was generated, D-optimal sampling schedules were derived, and D-optimal batch-sequential design was performed. We could show, for the same parameter sets, that for this protocol, two instead of three support points were already sufficient for full parameter identification under conditions where measurement noise was negligible. Importantly, under conditions in which measurement errors were not neglected, the number of replicates required to infer all parameters of the model up to a credibility interval of one hour was dramatically reduced (see Fig. 2.10).
In **Chapter 3**, we applied the mathematical framework, developed in Chapter 1, to improve our understanding of cell proliferation and migration in germinal centers. We first provided two rational

arguments, which, in line with former studies, suggest, that the classical notion of proliferation in germinal centers is difficult to reconcile with some of the previously published germinal center cell cycle data sets (see Section 3.2.1). We then developed and solved a mathematical model, in which germinal center B cells progress through the cell cycle completing the G_1, the S, and the G_2/M phase, with the additional fates however to enter a selection phase after the G_1 phase, and to leave the germinal center or re-enter the S phase after selection has been completed (see Fig. 3.3 A). By comparing the analytical predictions of this model with data from DNA-BrdU pulse-chase labeling experiments, we found that the model could explain the empirical kinetics best, when the cells divided in average every ten hours (see Fig. 3.3 B). To further investigate the interplay between migration and cell proliferation in germinal centers, we set up a second model, targeting migration and cell cycle progression specifically in the germinal center dark zone. In our hands, the most simple model which was able to describe simultaneously the considered data sets, comprised two distinct kind of cells, one trafficking and one self-renewing subpopulation (see Fig. 3.5). An analysis based on this model predicted a surprisingly high influx of cycling germinal center B cells into the dark zone. However more experimental and theoretical work is needed, in order to test the model assumptions in a single consistent system.

4.2 Discussion

This thesis provides a general mathematical tool to analyze DNA nucleoside analog pulse-chase labeling experiments. This tool has proven useful in the process of optimizing and testing *in silico* experimental pulse-chase labeling protocols, and showed itself versatile enough to describe nucleoside analog pulse-chase labeling experiments of exponentially growing and homeostatically regulated cell populations. Furthermore, it could be adapted to interpret additional types of data, for instance time-series data generated by stathmokinetic and photoactivation experiments. However, as with any other model-based approach, a set of assumptions had to be made in order to generate model predictions which permit comparison with experimental data. In the following, we will review some of our assumptions, discuss how much we depend on them, how they can be tested and to what extend they limit the applicability of our approach.

Model assumptions

The most basic assumption of our model, namely the sequential progression of proliferating cells through the various cell cycle phases, does probably not require any further justification. Overwhelming experimental evidence, collected over several decades of cell cycle research, has shown that this describes indeed very faithfully the way how mammalian cell division is accomplished (see Chapter 1, Motivation and Background). That the completion of each of the cell cycle phases takes a certain minimal amount of time, the latter being subject to stochastic variation, represents our second most basic assumption. In this general form, it is an almost trivial statement, given the ubiquitous thermodynamic fluctuations underlying biological processes, including DNA synthesis and gene expression [92, 173–175].
Throughout this thesis, we considered the delayed negative exponential distribution as a reasonable approximation to mimic phase completion times. This distribution was chosen as a trade off between mathematical convenience and biologically reasonable assumptions. It allowed us to derive analytical solutions for the kinetics of nucleoside analog pulse-chase labeling and similar experiments. In addition, this distribution seemed sufficiently flexible and accurate to reproduce well experimentally determined division time distributions and kinetic data from pulse-chase labeling experiments. Nevertheless, more realistic phase completion times might be considered in the future. The expressions presented in Chapter 1, especially those for the average kinetics of labeled and unlabeled cells expected from DNA nucleoside analog pulse-chase labeling experiments, are given as one-dimensional integrals and

convolutions, which can in principle be evaluated for more complex phase completion times on any modern computer. In addition, the fact that $n_0^{\phi}(t)$ (see Eq. 1.24) are derived as a function of the Laplace transform of the completion time density distribution implies that any probability density function with closed-form Laplace transform can be plugged into the expressions in order to yield analytical results for this specific population. Generalizing to the case of $n_i^{\phi}(t)$, for $i > 0$, is however hampered by the need to compute inverse Laplace transforms, which are typically hard to evaluate for non-standard density distributions. There exist however powerful and accurate numerical Laplace inversion algorithms [176], that are fast enough to perform even Bayesian inference in a reasonable amount of time.

Empirical phase completion times

One direct approach to measure phase completion time distributions, employs time lapse imaging. For this method, genetically modified cells 'communicate' their current cell cycle phase by means of cell cycle dependent fluorescent reporter genes [28, 72]. Currently, continuous imaging of single cells over more than 3 hours is however constrained to *in vitro* studies and relatively few cells. In addition, resolving all phases by microscopy is not possible in the FUCCI system [28], or requires to interpret subtle changes in the distribution of reporter genes inside the cytoplasm and the nucleus [72].
A second so far unexplored indirect approach could exploit the nucleoside analog double pulse-chase labeling technique proposed in Chapter 2 of this thesis. To see this possibility, consider for instance measuring $\mathrm{f}_{\mathrm{S,S}}^{lu}$ (i.e., the fraction of labeled undivided cells in S phase) at many different time points after the first pulse. Because, according to Eq. 1.20 and Eq. 1.24,

$$f_{\tau_S}(t) \propto e^{\mu t}\frac{d}{dt}e^{-\mu t}\frac{d}{dt}e^{\mu t} \times \mathrm{f}_{\mathrm{S,S}}^{lu}(t), \tag{4.1}$$

one finds that $f_{\tau_{\mathrm{S}}}$(the probability density function of τ_{S}) is uniquely defined by $\mathrm{f}_{\mathrm{S,S}}^{lu}$ and μ. Hence, the more accurately we measure $\mathrm{f}_{\mathrm{S,S}}^{lu}$ and μ (the growth rate μ can be estimated independently by cell counting or stathmokinetic experiments), the better will be our estimate of $f_{\tau_{\mathrm{S}}}$. By analogy, the same is true for $\mathrm{f}_{\mathrm{G_2M,G_2M}}^{u}$. The situation for the $\mathrm{G_1}$ phase is different. Using the double pulse labeling approach, the population $\mathrm{f}_{\mathrm{G_1,G_1}}^{u}$ cannot be measured as a pure population, because this population is still intermixed and indistinguishable from $\mathrm{f}_{\mathrm{G_2M,G_1}}^{u}$. Continuous labeling with the nucleoside analog from the first pulse, additional labeling with CFSE [39] or blocking cells in M phase by colchicine, could however prevent intermixing of $\mathrm{f}_{\mathrm{G_1,G_1}}^{u}$ with $\mathrm{f}_{\mathrm{G_2M,G_1}}^{u}$. Obviously, many support points will be required in order to generate accurate estimates, which makes this approach quite labor-intensive. Crucially however, the estimates will only depend on a single assumption, namely exponential growth, which can be tested independently by analyzing the growth curve. Interestingly, the same reasoning can also be applied to a homeostatically regulated population, in which the only assumption is the lack of growth (i.e., $\mu = 0$). In this case, Eq.4.1 simplifies, and we get $f_{\tau_\phi}(t) \propto \ddot{\mathrm{f}}_{\phi,\phi}(t)$, where the double dots indicate differentiating $\mathrm{f}_{\phi,\phi}(t)$ twice with respect to time.

Cell death

Another assumption, which was made in Chapter 1, is the absence of cell death. How much do the results and the methodology depend on this rather extreme condition? If cell death would happen for example shortly after mitosis, and as long as in average the number of cells that are born per unit of time is greater than the number of cells that die per unit of time, our formalism would apply without major modifications. Uniquely the factor two in the matrix Eq. 1.9 would have to become smaller. In the more realistic case that cells do not die immediately after mitosis, but instead at the restriction point [177], it would be reasonable to split the $\mathrm{G_1}$ phase into two parts, and introduce an additional parameter which specifies the fraction of cells that die at this point. Again, except of slight adaptations, most formulas derived in Chapter 1 would remain valid.

Ideally however, instead of making assumptions, one would like to assess cell death directly. Measuring when and to what extend BrdU positive cells appear after a first nucleoside analog pulse in the Annexin-V (a marker for apoptosis) positive compartment could provide the information necessary to infer the rate and duration of apoptosis *in vivo*. Surprisingly, we could not find any published study, which reports on or analyzes quantitatively such type of data.

Heterogeneity

In principle, the model-based inference approach presented herein can be applied to infer cell cycle progression in any homogeneous cell population[1], as long as cells can be identified based on cell surface markers, tissue, or other characteristics. If the population under study is however heterogeneous, meaning that it is composed of a set of phenotypically indistinguishable subpopulations with potentially different cell cycle progression parameters, the analysis becomes obviously more complex. A good example is given by the germinal center dark zone B cell population, analyzed in Chapter 3. While data from photoactivation experiments indicate heterogeneity in respect to migration in single lymph node GCs, an aspect which we tentatively incorporated into our analysis, further subpopulations might have remained undetected. For instance, we completely ignored B cell receptor affinity dependent proliferation, or differences in cell cycle progression due to local conditions in specific GCs. While the existence of undetected subpopulations can never completely be ruled out given the indirect nature of our analysis, there are some read-outs from nucleoside analog pulse-chase labeling experiments which can at least serve as an indicator whether the assumption of homogeneity in respect to cell cycle progression represents a reasonable approximation. First of all, the kinetics of labeled and unlabeled cell populations in the various phases clearly depend on the composition of the population. Imagine, for example, that some cells in a given population of interest require substantially more time to complete the G_2M phase compared to their peers. This would become apparent in the nucleoside pulse-chase labeling data, as the evolution of $\mathrm{f}^u_{\mathrm{G_2M,G_2M}}$ would not correspond to the kinetics expected from a homogeneous model. Therefore, if in practice a homogeneous model performs well in reproducing the data, it is not completely unreasonable to assume that the population is indeed homogeneous. In contrast, if a homogeneous model fails to reproduce the data, one may consider the presence of heterogeneous subpopulations as one possible cause for the failure. From a more formal perspective, we may recall that $f_{\tau_{\mathrm{G_2M}}}$ is fully defined by $\mathrm{f}^u_{\mathrm{G_2M,G_2M}}$, if the population size remains constant over time. In our example with two subpopulations, and assuming in addition homeostatic regulation, $\mathrm{f}^u_{\mathrm{G_2M,G_2M}}$ will correspond to the weighted sum of the average kinetics from the fast and the slow progressing subpopulation. As $f_{\tau_{\mathrm{G_2M}}}$ can be derived from $\mathrm{f}^u_{\mathrm{G_2M,G_2M}}$ by differentiating twice (see above), $f_{\tau_{\mathrm{G_2M}}}$ will also correspond to a weighted sum or a mixture of two completion time density functions. The latter could be analyzed using for example a mixture model in order to tease apart both density functions. Besides of testing, whether experimentally determined kinetics are consistent (or not) with a homogeneous model, it is also possible to check independently whether S phase durations are homogeneous or heterogeneous. Given that nucleoside analog incorporation is proportional to the DNA synthesis rate, and that DNA synthesis rate is proportional to the S phase duration, similar amounts of incorporated BrdU indicate similar S phase durations. The four BrdU pulse-chase labeling data sets that we analyzed in this thesis, showed quite narrow uni-modal distributions of BrdU incorporation, suggesting in our case approximately homogeneous populations in respect to S phase progression.

Bayesian inference

Bayesian inference has played several distinct and important roles in the course of this thesis. In Chapter 1, it was key in discovering and exploring the uncertainty in the parameter estimates inferred from the U87 and the V79 data sets. In Chapter 2, it formed the basis for optimal experimental

[1] Notice that even though the population is assumed homogeneous, the completion times are subject to stochastic variation and are therefore in some sense heterogeneous.

design, as the uncertainty in the parameter values, determined through the Bayesian approach, was the quantity that we aimed to minimize. Finally in Chapter 3, where due to additional phases and a second cell type, the number of parameters increased significantly, Bayesian inference helped to identify the quantities which, despite the increasing number of degrees of freedom, remained relatively well defined by the data (e.g., the average transit time, the average S phase duration).
It should however be emphasized that the accuracy of Bayesian inference depends on the accuracy of the underlying noise model. In this thesis, we have chosen a very simplistic noise model, namely a continuous approximation of a scaled multinomial distribution, involving a single additional parameter (see 1.5.2). This choice appeared more appropriate to describe fluctuations in frequencies, compared to the more commonly used beta or normal densities, yet the data that we analyzed was to sparse, in order to test accurately, how well this model reproduces the actual correlation structure in the measured frequencies. Because the measurement errors are most likely composed of contributions from a range of different sources, for instance, variability in experimental conditions, gating errors, stochasticity in cell division, FACS measurement errors, inter-animal heterogeneity, the true error distribution is probably experiment-specific and more complex than our *ad hoc* assumptions. More extensive data sets are therefore needed to define a proper and hopefully general noise model. In the mean time, we propose to consider Bayesian inference, in the form presented here, as a useful qualitative rather than an accurate quantitative tool to understand uncertainty in cell cycle parameter estimates. To the best of our knowledge, no study has been published so far which applied Bayesian inference to cell cycle parameter estimation. The same is also true for rational experimental design.

4.3 Outlook

The future of nucleoside analogs

Nucleoside analogs have now been used by biologists for over seventy years to study cell proliferation. And, as pointed out in a recent review by Cavanagh et al. [13], BrdU alone has been utilized in over 20,000 biomedical studies to track newly synthesized DNA. This makes it unlikely that the use of thymidine or more generally nucleoside analogs will be discontinued in the near future. On the contrary, the discovery of EdU has provoked renewed interest from the biological community in the potential of this technique. Currently however, many studies refrain from measuring DNA content concurrently with nucleoside incorporation, much less consider to use model-based analysis or administration of several pulses in order to increase the accuracy of the estimates.
Despite their popularity among biologists, a specific aspect of nucleoside analogs prevents their widespread use in the clinics. Because analogs slightly differ from the originals, their incorporation into newly formed DNA can induce DNA instability and increase the probability for malicious mutations to occur. While several studies, where BrdU has been administrated to humans have been approved in the past [20,33,178–181], patients enrolled in these experiments typically suffered from diseases like cancer and HIV, with notoriously bad long-term prognosis. Administration of nucleoside analogs to potentially healthy individuals is clearly more difficult to justify, even if their toxic effects, similar to e.g., radiation therapy, are most likely a question of dosage. To avoid these issues, non-toxic nucleoside analogs which contain the stable isotopes ^{2}H, ^{13}C and ^{15}N, have been explored in several studies [64, 182]. Unfortunately, these are, in contrast to fluorescently labeled BrdU, currently impossible to detect at a single-cell level [183].
Irrespectively, inferring precise cell cycle kinetics from tumor tissues on a patient-to-patient basis remains an attractive notion which could support individualized cancer therapy in the future. For this, the triple pulse labeling method, discussed at the end of Chapter 2, applied either with conventional or so far undiscovered non-toxic nucleoside analogs, might prove useful, as a single tissue sample (biopsy) may already contain sufficient information to identify cell cycle parameters, possibly relevant for the prediction of tumor growth and susceptibility to certain drugs.

Estimating variability in cell cycle progression : Hypothetical applications for tumor growth prediction and treatment?

Considerable amount of work has been invested in the past, to characterize cell cycle progression in tumors (e.g., [8, 20, 30, 31, 35, 81, 83, 84, 125, 178, 184–186]). This impressive effort was driven by the hope that some properties of cell cycle progression in cancerous cells would be correlated with disease stages or treatment outcomes. Most prominent quantities that have been considered are the potential doubling time and the average duration of the S phase, although their relevance is not undisputed [8, 187]. Interestingly, despite the fact that some cancerous cells are known for their heterogeneity in respect to cell cycle progression (see e.g., Table 2.1 and [21,188]), apparent for instance through a very broad second wave in FLM experiments, this specific aspect has received relative little attention in applications in the clinics, perhaps because variability has been difficult to infer with currently available techniques. In this section, we will discuss two examples which illustrate how variability in cell cycle progression may play a role in tumor growth prediction and treatment. These examples are meant as simple and naive toy models, as they do not account for the high complexity encountered in real-life cancer disease.

For the sake of simplicity, we will consider a cell cycle model with a single phase (instead of three as in Chapter 1). The completion time for this phase is assumed, as before, shifted negative exponentially distributed. This corresponds one-to-one to the classical Smith-Martin cell cycle model [12]. In the absence of cell death, the asymptotic growth rate μ of an asynchronously dividing cell population which follows this model is given by the solution of

$$\gamma = \mu,$$

where γ is defined by Eq. 1.8. One finds that

$$\mu = -\frac{1}{\alpha} + \frac{W(\frac{2\beta e^{\beta/\alpha}}{\alpha})}{\beta}, \tag{4.2}$$

where $W(\cdot)$ is the so-called Lambert W-function. This expression approaches $1/\alpha$ for $\beta \to 0$, and $\ln 2/\beta$ for $\alpha \to 0$, which is expected (see Section 1.3). Because the average division time for this model is $\alpha + \beta$, Eq. 4.2 shows that the growth rate of the population in this model is not simply a function of the average division time, but does depend in a more subtle way on both parameters α and β.

For the sake of the argument, imagine that we have measured, from e.g., a biopsy, the average division time for a non-growing but proliferating tumor cell population to be 12 hours (for instance by the relative movement method [15]), but that the data or the method did not allow us to estimate the variability in cell cycle progression. Furthermore, suppose, that we aim at using the information contained in our data to predict re-growth of the tumor immediately after surgical removal of neoplastic tissue, under the simplifying assumptions that cell cycle parameters of residual cancer cells remain unchanged and that cell death becomes negligible shortly after surgery. The motivation for such an approach is that these predictions could inform treatment design, as for example patients with faster growing tumors might, due to a more rapid repopulating of remaining cancer cells, require, in order to become effective, shorter intervals between subsequent radio- or chemo-therapies [187].

However, from a theoretical perspective, predicting population growth based on the average division time alone is virtually impossible. To see this, we plotted in Fig. 4.1 the expected sizes of two populations starting both with a single cell, a first population growing over 3 days with rate $\mu = 1/12$ and a second population growing during the same time at a lower rate, $\mu = \ln 2/12$ respectively. These are two cases which, under our model, are perfectly compatible with an average division time of 12 hours ($\alpha = 12$, $\beta = 0$, $\mu = 1/12$, or $\alpha = 0$, $\beta = 12$, $\mu = \ln 2/12$). The graph and a simple computation show that the size of the first population becomes more than 9 times larger (i.e., $\exp(\frac{1-\ln(2)}{12} \times 24 \times 3) = 9.1$) after only 3 days, compared to the slower growing 'deterministic'

population, even though the average division time is 12 hours for both.
For the second example, consider again the single-phase cell cycle model as specified before. This time, we are interested in estimating the minimal time, it would take an hypothetical 100% effective drug, which exclusively acts during cytokinesis, to kill in 99% of the cases all cycling cells of a tumor population. Suppose that at the time of treatment the size of the cycling cell population in the tumor is $n = 10000$. Assume furthermore, that we have measured as before the average division time $\alpha+\beta = 12$ hours, and that we want to derive the appropriate treatment duration d. For the unrealistic case that $\alpha = 0$, the time it will take all n cells to enter cytokinesis will obviously be $\beta = 12$ hours. For the other extreme, where $\alpha = 12$ hours and $\beta = 0$, estimating the time at which the last of the n cells will enter cytokinesis is slightly more intricate. We know that the probability that n iid random variables are smaller than a given value x equals $F(x)^n$, where $F(\cdot)$ is the cumulative distribution of the random variables, which in our case is the division time. For negative exponentially distributed division times with mean α and $F(x) = 1 - e^{-x/\alpha}$, we thus compute, exploiting the Markov property, the probability for a successful treatment (i.e., which kills all the cells) with duration d as follows

$$\mathbb{P}(\text{success}) = (1 - e^{-d/\alpha})^n. \tag{4.3}$$

Because we require for our treatment design this probability to become 0.99, we get

$$d = -\alpha \times \ln(1 - 0.99^{1/n}). \tag{4.4}$$

Generalizing this expression for the case where $\beta > 0$, the minimal treatment length becomes instead,[2]

$$d^* = -\alpha \times \ln(1 - 0.99^{1/n^*}) + \beta, \tag{4.5}$$

where $n^* = n \times (2 - e^{\beta\mu})$. Fig. 4.1 shows d^* as a function of the parameter α, given a fixed average division time of 12 hours. This shows that the treatment time, as defined above, increases almost linearly with α, and becomes approximately 7 days for $\alpha = 12$ hours.
Both examples highlight how variability in division times, irrespective of their average durations, can have a significant influence on the evolution of dividing cell populations in the presence and in the absence of cytotoxic drugs. It may therefore become informative, in certain cases, to assess, in addition to the classical average quantities, variability in cell cycle progression in order to predict and understand tumor growth or the effectiveness of anti-cancer drugs. In this thesis, we presented a methodology to estimate with high precision and accuracy the variability in cell cycle phase completion times, which at the same time allows to derive, assuming independency of completion times, the variability in division times.

[2]The derivation is as follows: Assume, as proposed in the original Smith and Martin model, that the cell cycle is divided into an A state with exponential completion times and a B state with deterministic completion times. It can be shown that the proportion of cells in the A state is given by $2 - e^{\beta\mu}$. The probability that all cells in the A state (there are in average $n^* = n \times (2 - e^{\beta\mu})$), leave the A state x hours after treatment initiation is $(1 - e^{-x/\alpha})^{n^*}$. Therefore, if our drug would act in the B state, the target treatment duration would become $-\alpha \times \ln(1 - 0.99^{1/n^*})$. However, because by assumption our drug acts during cytokinesis, the last cell that leaves the A state still has to complete the B state in order to be killed, and therefore we get for the treatment duration $d^* = -\alpha \times \ln(1 - 0.99^{1/n^*}) + \beta$.

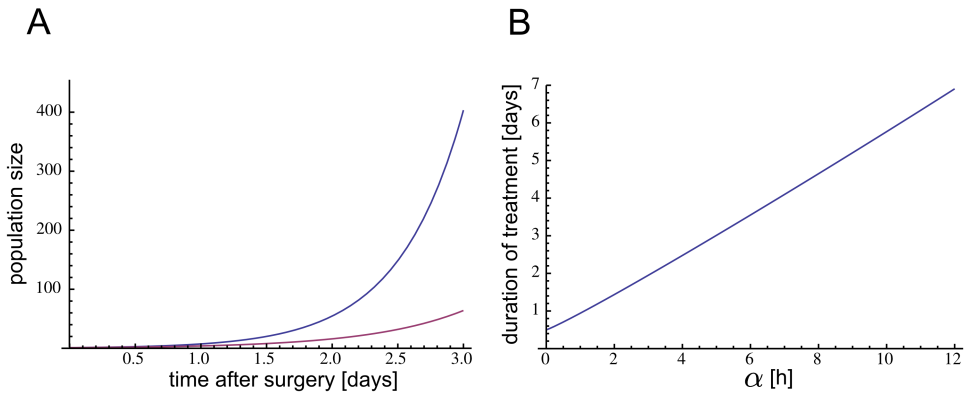

Figure 4.1: Impact of variability in division times on population kinetics. **A)** Sizes of two populations, starting with a single cell and proliferating over three days, one with purely exponential distributed division times, a second with fixed (deterministic) division times, and both with the same average division time of 12 hours. This shows that after three days, the faster growing population (exponential) is about 9 times larger than the slower growing deterministic population. **B)** Treatment length d^* as a function of α, required to kill, in 99% of the cases, all cycling cells in a tumor population (see main text for model assumptions). While for $\alpha = 0$ (zero variability), the treatment length is 12 hours, for $\alpha = 12$ hours (maximal variability) the required treatment length becomes almost as long as a week. The average division time is assumed to be 12 hours, independently of α.

Algorithms

Several classical and more recent algorithms were used in this thesis to analyze the models and data sets at hand. Most of these algorithms are well described in the literature, and shall therefore be covered in this section only very briefly. We will also restrict our attention to those algorithms which have been implemented by the author or were used in form of open source libraries as part of other C++ programs. For example, the Risch algorithm, which forms the basis of most symbolic integration routines (including Mathematica's), will not be described, although without this algorithm the majority of theoretical results presented in this thesis, would have been very hard to derive. The interested and brave reader is refered to [189].

The Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm is a well-studied algorithm, commonly employed by statisticians to derive maximum-likelihood estimates for models with latent variables [121]. The basic idea behind the algorithm is to use current parameter estimates and available data to derive in a first step (E-step) the expected log-likelihood or the Q-function, with respect to the conditional distribution of the latent variables. In a second step (M-step), the parameters are updated such that the Q-function is maximized. Both steps are iterated until some convergence criteria are met.
In Chapter 2 of this thesis, we use the EM algorithm to estimate the density of D-optimal designs in a three-dimensional design-space, based on a cloud of N previously derived D-optimal designs for specific parameters. The density is assumed to be accurately approximated by a multivariate gaussian mixture model with $K = 10$ kernels. For the kth kernel with weight α_k and dimension $d = 3$ the density is defined by

$$p_k(x|\theta_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)},$$

where $\theta_k = \{\mu_k, \Sigma_k\}$ are the parameters, i.e., the mean and the covariance matrix, which typically specify a multivariate Gaussian density, $|\Sigma_k|$ is the determinant of the covariance matrix and $(x-\mu_k)^t$ is the transpose of the column vector $(x-\mu_k)$, which is the difference vector between the mean vector μ_k and a given position vector (or a data point) x.
The gaussian mixture model is then given by a weighted sum over the K kernels, i.e.,

$$p(x|\theta) = \sum_{m=1}^{K} \alpha_m p_m(x|\theta_m).$$

and the latent variables are for each D-optimal design the so-called 'membership weights', which correspond to the probabilities a given design belongs to a given kernel.
The EM update equation for gaussian mixture models, given the initial estimates α_k and θ_k are

- E-step:

$$\omega_{ik} = \frac{p_k^{(i)}(x_i|\theta_k)\,\alpha_k}{\sum_{m=1}^{K} p_m(x_i|\theta_m)\,\alpha_m}, 1 \le i \le N \text{ and } 1 \le k \le K$$

- M-step

$$\alpha_k^{(i+1)} = \frac{N_k}{N},$$
$$\mu_k^{(i+1)} = (\frac{1}{N_k}) \sum_{i=1}^{N} \omega_{ik} x_i$$
$$\Sigma_k^{(i+1)} = (\frac{1}{N_k}) \sum_{i=1}^{N} \omega_{ik} (x_i - \mu_k^{(i+1)})(x_i - \mu_k^{(i+1)})^t.$$

where $N_k = \Sigma_{i=0}^{N} \omega_{ik}$.
For the implementation of the EM algorithm, the well structured Matlab code from [190] was translated into C++ code, using one of the fastest open source linear algebra libraries, called Eigen (http://eigen.tuxfamily.org). Several details in the algorithm proposed by Figueiredo *et* al. differ from the traditional implementation. First, kernels are updated one at a time and not concurrently, presumably to avoid convergence to local maxima. Secondly, kernels are deleted from the mixture, using as a criterion the minimal description length. For estimating the density in the design space this feature was disabled, as an optimal number of kernels was not required. The routine implemented in C++ can easily handle several thousand data points, and was successfully used, besides of estimating the density in design space, to cluster FACS data with more than 3×10^5 single measurments (for an example with less data points see Fig. 2). In addition, a faster version was developed, which is based on binned data (e.g., pixmaps), instead of continuous data points.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm represents one of the most verstile Marcov chain Monte Carlo (MCMC) methods to generate sequences of random samples from a multivariate probability distribution for which direct sampling is not straightforward [191]. It was first proposed in a paper by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller [192] and was further generalized by Hastings [193].
In order to generate a sample from the distribution $\pi(\cdot)$ using the MH algorithm, a suitable proposal density or candidate-generating function $q(x, y)$ has to be specified. The latter is usually a distribution from which it is easy to sample from, and defines, conditioned on the position x, the probability to draw as a next position y. For the special case that $q(x, y) = q(y, x)$, the basic steps of the MH algorithm are, starting at $x^{(1)}$:

- repeat for $i = 1, 2, \ldots, N$
- draw y from $q(x^{(i)}, \cdot)$, and u from the uniform distribution $\mathcal{U}(0, 1)$
- if $u \leq \frac{\pi(y)}{\pi(x)}$ then set $x^{(i+1)} = y$
- else set $x^{(i+1)} = x^{(i)}$
- return the values $\{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$

It can be shown that this sequence converges for large N to a sample from $\pi(\cdot)$.
While represting an extremely powerful algorithm, a serious issue with MH, especially for high dimenional problems, can be the rate of convergence, which depends both on $\pi(\cdot)$ and $q(\cdot, \cdot)$. As a general rule, it is found that convergence is fastest if $\pi(\cdot) \approx q(\cdot, \cdot)$, however choosing $q(\cdot, \cdot) = \pi(\cdot)$ is obviously not practical. Therefore several authors [91, 194, 195] have proposed adaptive variants of the MH algorithm, where $q(\cdot, \cdot)$ is updated as soon as more information about $\pi(\cdot)$ is available. Under relatively mild conditions, convergence is guaranteed even if $q(\cdot, \cdot)$ changes over time [196].

For this thesis, the M-H algorithm proposed by Roberts *et* al. [91] was implemented, in which a d-variate normal proposal density is updated in the i-th step (for $i > 2d$) as follows

$$q^{(i)}(x, \cdot) = (1 - \beta)\mathcal{N}(x, 2.38^2\Sigma^{(i)}/d) + \beta\mathcal{N}(x, (0.1)^2 I_d/d),$$

where $\Sigma^{(i)}$ is the current empirical estimate of the covariance structure of the target distribution based on the run so far, and where β is a small positive constant. The constants 2.38 and 0.1 are empirical values which under certain conditions optimize mixing of the Marcov chain.
It was observed that mixing of the chains was generally satisfactory, despite the high dimensions and the extreme correlations frequently encountered in the posteriors over the cell cycle parameters.

The Covariance Matrix Adaptation Evolutionary Strategy

The covariance matrix adaptation evolutionary strategy (CMA-ES) is a stochastic population based method for parameter optimization of non-linear functions. It uses processes inspired by Darvinian evolution, like reproduction, mutation, and selection in order to find local but also global maxima. In contrast to genetic algorithms, where the domain of the functions to be optimized is often dicrete, CMA-ES requires continuous parameters which are 'mutated' by sampling from a multivariate normal distribution. The equation for the positions of the offsprings at generation $i + 1$ is, according to [197]

$$x_k^{(i+1)} \sim \mu^{(i)} + \sigma^{(i)}\mathcal{N}(0, \Sigma^{(i)}) \quad \text{for } k = 1, \ldots, n \tag{6}$$

where

$\sim$ denotes the same distribution on the left and on the right hand side,

$\mathcal{N}(0, \Sigma^{(i)})$ is a multivariate normal distribution with zero mean and covariance $\Sigma^{(i)}$,

$x_k^{(i+1)}$ is the position of the k-th offspring at generation $i + 1$,

$\mu^{(i)}$ is the mean positions of the search distribution at generation i,

$\sigma^{(i)}$ is the overall standard deviation at generation i,

$\Sigma^{(i)}$ is the covariance matrix at generation i,

n is the offspring population size.

The average position of the search distribution at the $(i + 1)$-th generation is given by

$$\mu^{(i+1)} = \sum_{j=1}^{m} \omega_j x_{j:\lambda}^{(i+1)} \quad \text{with} \quad \sum_{j=1}^{m} \omega_j = 1$$

where

$m \leq n$ is the parent population size,

$\omega_{j=1\ldots m}$ are positive weight coefficients for recombination,

$x_{j:\lambda}^{(i+1)}$ are the j-th best solutions from Eq. 6.

The selection mechanism corresponds to *truncated selection* by choosing m out of n offsprings which contribute to the average position of the search distribution at the next generation.

The update equation for the covariance matrix is more intricate

$$\Sigma^{(i+1)} = (1 - c_{cov})\Sigma^{(i)} + \frac{c_{cov}}{m_{cov}} p_c^{(i+1)} p_c^{(i+1)^T} + c_{cov}(1 - \frac{1}{m_{cov}}) \times \sum_{j=1}^{m} \omega_j y_{j:n}^{(i+1)} (y_{j:n}^{(i+1)})^T, \tag{7}$$

where

$$p_c^{(i+1)} = (1 - c_c) p_c^{(i)} + \sqrt{c_c(2 - c_c) m_{eff}} \frac{\mu^{i+1} - \mu^i}{\sigma^i}$$

is the so-called evolution path, and $y_{j:n}^{(i+1)} = (x_{j:n}^{(i+1)} - \mu^i)/\sigma^i$.
The parameters m_{cov}, c_{cov}, m_{eff} and c_c are tuning parameters which depend on the dimensionality of the problem and the desired stringency of selection. They are mostly determined based on empirical experience. An in-depth description of each of the terms in Eq. 7 is beyond the scope of this section and can be found together with the update equation for $\sigma^{(i+1)}$ in [118, 197].
The CMA-ES algorithm is implemented in the open-source library shark ([122]) which can be readily included into any application written in C++. In this thesis, the algorithm was successfully used for constrained non-linear least-squares-fitting of model predictions with six and up to 18 parameters. In Chapter 2, the routine also served to find the D-optimal and non-local optimal designs in up to six dimensions. Finally it has also been incorporated into the web-based application 'Cell Cycle Timer' (see Appendix, Software).

Software

In the context of this thesis, several software applications with graphical user interfaces were developed to assist in the data analysis, to explore classical and novel algorithms, and to make some of the results derived in Chapter 1 and Chapter 2 more easily available to a broader audience.

GEM xD

GEM xD, which is short for **G**ating with the **E**xpectation-**M**aximization Algorithm in any dimension, is an interactive 3-D FACS data viewer, which allows to cluster high-dimensional FACS data using a fast and efficient implementation of the expectation-maximization (EM) algorithm (see Appendix, Algorithm and [190]) specific for gaussian mixture models (GMM). The clustering is implemented as follows:

1. a GMM with a relative large number of kernels (circ. 10-30) is initialized with the standard k-means or the k-means++ algorithm [198].

2. the EM algorithm is employed to find maximum-likelihood estimates for the kernels.

3. the kernels of the GMM are then grouped according to basins of attraction.

4. a majority rule is used to assign data points *via* the kernels to the basins of attraction.

Notice that the final number of clusters (populations) corresponds to the detected number of basins of attraction.
EM in combination with GMM or similar models have previously been used to analyze FACS data, where a common approach is to assume that the number of kernels roughly matches the expected number of subpopulations in the total cell population [199, 200]. This works well if the subpopulations are distributed approximately like multivariate gaussian or similar distribution, which is however rarely the case. The method described above, by using far more kernels, achieves a better fit to the data. This allows to use the resulting GMM to identify the basins of attraction, which in turn leads to the definition of the subpopulations. A similar approach, however based on kernel density estimation, has been published recently by Ge *et* al. [201].
The interface of GEM xD is highly interactive, giving the user the possibility to move, rotate and zoom in 3-D the projection of the possibly high-dimensional data. Using the low-level rendering library openGL (opengl.org) in combination with Qt4 (qt-project.org) and C++ permits to view and manipulate large data sets ($> 10^5$) without significant performance issues. Gating can be done either by drawing a polygon on a 2-D plane in the 3-D space or more inventively by using the clusters obtained from the GMM. Clusters can be included or removed from the data based on their color. The utility of this approach is shown in Fig. 2, where six-dimensional experimental dual pulse data (shown are the BrdU-DAPI-EdU axis) is first initialized with the k-means algorithm (B), then clustered with a variant of the EM algorithm (C), and finally 'cleaned' by removing clusters apparently originating from doublets or other sources of noise (D).
GEM xD has extensively been used to gate the DAPI-BrdU single pulse labeling data analyzed in Chapter 1 and Chapter 3 of this thesis.

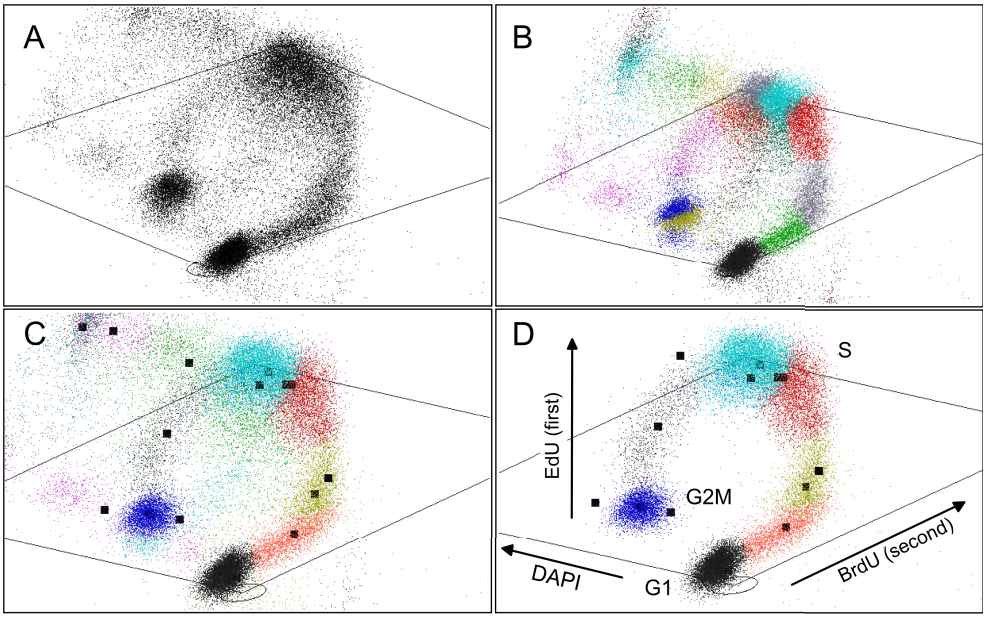

Figure 2: Typical work flow of GEM xD. First, high-dimensional FACS data is loaded (A). Then k-means clustering is applied to initialize a GMM (B). The EM algorithm is used to find the maximum likelihood of the GMM given the data (C). Populations that are identified by the user as noise are removed by selecting respective clusters based on their color (D). Finally the frequencies of the populations of interest are determined. Black squares represent local maxima. The data was generated, as a preliminary test for EdU-BrdU dual-pulse labeling experiments, by Telma Lopez and Rui Gardner at the Instituto Gulbenkian de Ciencia, Portugal.

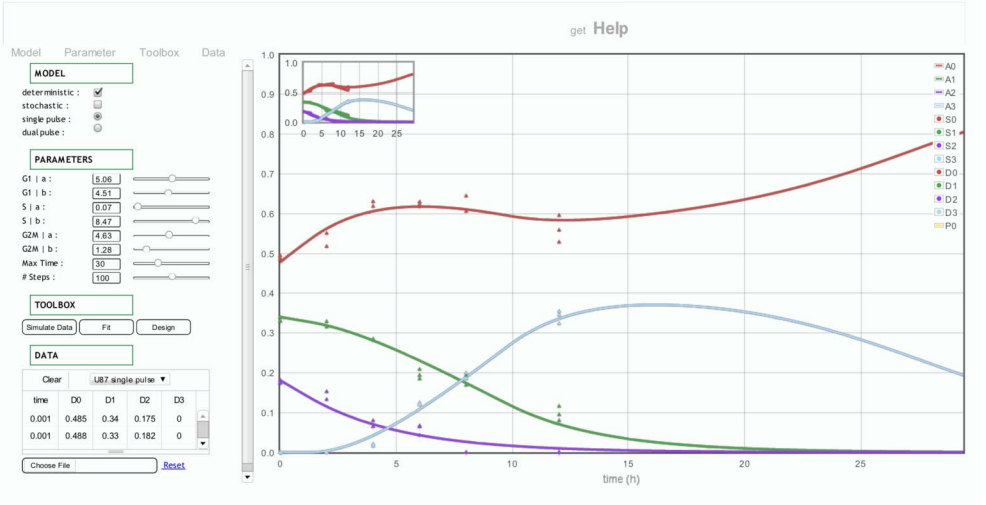

Figure 3: Graphical user interface for Cell Cycle Timer. On the left, parameters can be adjusted by moving the sliders, and options for e.g., the fitting and experimental design routines, can be specified. To the right, the U87 example data set (triangles) has been loaded and subsequently fitted with the model predictions (lines). Fitting with the CMA algorithm takes about 5 seconds on a normal computer.

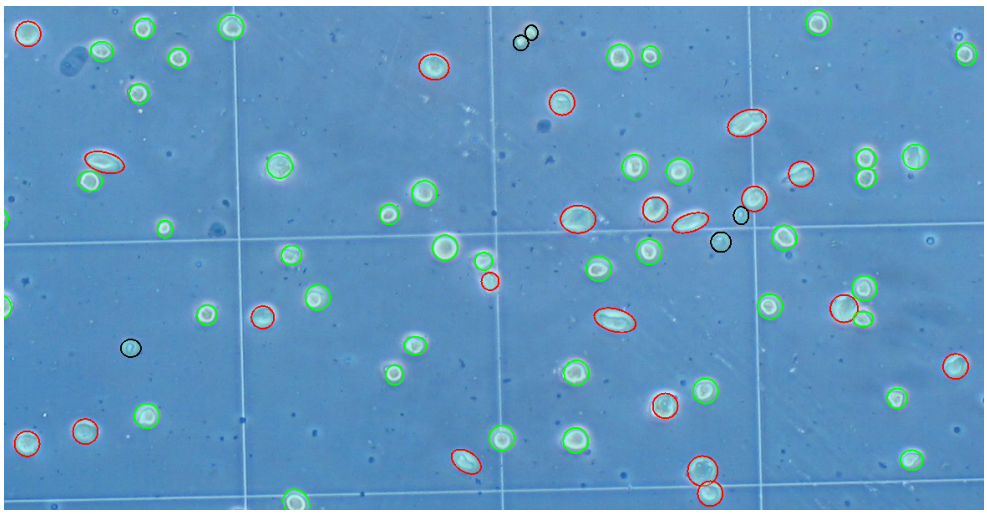

Figure 4: Cell Qounter was used to manually count three types of cells (green:normal; red:dead; black:unclassified) on a light-microscopy image from the human glioma cell line U87. The image was recorded by Irene Jaehnert at the Department of Neurosurgery, Ludwig-Maximilians University Munich, Klinikum Grosshadern.

Cell Cycle Timer

Cell Cycle Timer (cellcycletimer.appspot.com) is an intuitive, graphical web-based application, whose target audience are experimentalist, interested in including design of experiments (as developed in Chapter 2) into their proliferation studies (see Fig. 3). Its main features are:

- parameter exploration of the models presented in Chapter 1 and Chapter 2
- stochastic simulations to visualize the impact of noise on the kinetics
- generation of synthetic data
- loading of user data
- example data (e.g., the U87 and V79 data sets)
- fitting of the single and dual pulse model to data using the CMA algorithm
- experimental design (preliminary)

Experimental design, although being the main motivation to implement this application, is still under development and is currently based on a full batch-sequential design algorithm to decide for the next optimal support point.
A peculiarity of Cell Cycle Timer is that it is an application based simultaneously on three programing languages, namely Html5, Javascript and C++. Computationally intensive tasks, like model fitting and experimental design are implemented in C++ and run on the client side in a so-called sandbox, the interactive graphical user-interface relies on Javascript and finally for structuring and presentation of the application's content and for loading and saving files, Html5 is utilized. This unusual combination is made possible by a recently developed technology named 'native client' (chromium.org/nativeclient), which allows to run compiled C++ code inside of the chromium web browser. Users which have the web browser chromium or chrome installed (currently, about 40% of all users in the Internet) can (disabling some security features) use the application as it it were a normal web page.

Cell Qounter

Cell Qounter is a simple Qt based application to count efficiently cells (or other objects) on light-microscopy or other 2-D images. The user can, besides of basic features like loading images, click and drag interactivity, zoom in and zoom out, mark a cell by placing a resizable and movable semi-transparent ellipsoid on top of it (identification). If cells of different types are to be counted (e.g. living and dead cells) this information can be easily added (classification). By selecting a given ellipsoid and pressing a key (e.g., 1-3) on the keyboard, the ellipsoid changes its color accordingly. Because identification is most efficiently done at a certain 'distance' from the image, while accurate classification requires a close-up view, Cell Qounter allows to first identify the cells at a certain magnification and then swap quickly through all the ellipsoids using the arrow keys on the keyboard, automatically clipping and magnifying the part of the image below the ellipsoid, i.e., the cell which is to be classified. The number of ellipsoid and their color are automatically recorded and can be retrieved from a dialog. Finally the positions and the color of the ellipsoids can be corrected at any time and saved together with the image file, to resume counting at a later time.
Several open source application to manually count cells are available, for example a plug-in for imageJ or the Python-based cellprofiler (cellprofiler.org). However, at the time of development, the simple task of manually counting cells seemed, given the above mentioned features, more practical and convenient with Cell Qounter, when compared to the available open-source alternatives.

Bibliography

[1] Massagué, J. (2004) G1 cell-cycle control and cancer. *Nature* **432**, 298–306.

[2] Lodish, H, Berk, A, Kaiser, C. A, Krieger, M, Scott, M. P, Bretscher, A, Ploegh, H, & Matsudaira, P. (2007) *Molecular Cell Biology.* (W. H. Freeman), 6th edition.

[3] Nishitani, H & Lygerou, Z. (2002) Control of DNA replication licensing in a cell cycle. *Genes to cells : devoted to molecular & cellular mechanisms* **7**, 523–534.

[4] Jallepalli, P. V & Lengauer, C. (2001) Chromosome segregation and cancer: cutting through the mystery. *Nature reviews. Cancer* **1**, 109–117.

[5] Scholzen, T & Gerdes, J. (2000) The Ki-67 protein: from the known and the unknown. *Journal of cellular physiology* **182**, 311–322.

[6] Stark, G. R & Taylor, W. R. (2004) Analyzing the G2/M checkpoint. *Methods in molecular biology (Clifton, N.J.)* **280**, 51–82.

[7] Li, F, Ambrosini, G, Chu, E. Y, Plescia, J, Tognin, S, Marchisio, P. C, & Altieri, D. C. (1998) Control of apoptosis and mitotic spindle checkpoint by survivin. *Nature* **396**, 580–584.

[8] Haustermans, K, Fowler, J, Geboes, K, Christiaens, M. R, Lerut, A, & van der Schueren, E. (1998) Relationship between potential doubling time (Tpot), labeling index and duration of DNA synthesis in 60 esophageal and 35 breast tumors: Is it worthwhile to measure Tpot? *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **46**, 157–167.

[9] Macdonald, P. D. M. (1970) Statistical Inference from the Fraction Labelled Mitoses Curve. *Biometrika* **57**, 489–503.

[10] Schultze, B, Kellerer, A. M, & Maurer, W. (1979) Transit times through the cycle phases of jejunal crypt cells of the mouse. Analysis in terms of the mean values and the variances. *Cell and Tissue Kinetics* **12**, 347–359.

[11] Takahashi, M. (1966) Theoretical basis for cell cycle analysis I. Labelled mitosis wave method. *Journal of Theoretical Biology* **13**, 202–211.

[12] Smith, J. A & Martin, L. (1973) Do Cells Cycle? *Proceedings of the National Academy of Sciences* **70**, 1263–1267.

[13] Cavanagh, B. L, Walker, T, Norazit, A, & Meedeniya, A. C. B. (2011) Thymidine Analogues for Tracking DNA Synthesis. *Molecules* **16**, 7980–7993.

[14] White, R. A, Terry, N. H, & Meistrich, M. L. (1990) New methods for calculating kinetic properties of cells in vitro using pulse labelling with bromodeoxyuridine. *Cell and Tissue Kinetics* **23**, 561–573.

[15] Begg, A. C, McNally, N. J, Shrieve, D. C, & Kärcher, H. (1985) A method to measure the duration of DNA synthesis and the potential doubling time from a single sample. *Cytometry* **6**, 620–626.

[16] Dolbeare, F, Gratzner, H, Pallavicini, M. G, & Gray, J. W. (1983) Flow cytometric measurement of total DNA content and incorporated bromodeoxyuridine. *Proceedings of the National Academy of Sciences of the United States of America* **80**, 5573–5577.

[17] Terry, N. H & White, R. A. (2006) Flow cytometry after bromodeoxyuridine labeling to measure S and G2+M phase durations plus doubling times in vitro and in vivo. *Nature protocols* **1**, 859–869.

[18] Yanagisawa, M, Dolbeare, F, Todoroki, T, & Gray, J. W. (1985) Cell cycle analysis using numerical simulation of bivariate DNA/bromodeoxyuridine distributions. *Cytometry* **6**, 550–562.

[19] Sherer, E, Tocce, E, Hannemann, R. E, Rundell, A. E, & Ramkrishna, D. (2008) Identification of age-structured models: Cell cycle phase transitions. *Biotechnol. Bioeng.* **99**, 960–974.

[20] Ritter, M. A, Fowler, J. F, Kim, Y. J, Gilchrist, K. W, Morrissey, L. W, & Kinsella, T. J. (1994) Tumor cell kinetics using two labels and flow cytometry. *Cytometry* **16**, 49–58.

[21] Baisch, H, Otto, U, Hatje, U, & Fack, H. (1995) Heterogeneous cell kinetics in tumors analyzed with a simulation model for bromodeoxyuridine single and multiple labeling. *Cytometry* **21**, 52–61.

[22] Schultze, B, Maurer, W, & Hagenbusch, H. (1976) A two emulsion autoradiographic technique and the discrimination of the three different types of labelling after double labelling with 3H- and 14C-thymidine. *Cell and tissue kinetics* **9**, 245–255.

[23] Hawkins, E. D, Turner, M. L, Dowling, M. R, van Gend, C, & Hodgkin, P. D. (2007) A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proceedings of the National Academy of Sciences* **104**, 5032–5037.

[24] Hauser, A. E, Junt, T, Mempel, T. R, Sneddon, M. W, Kleinstein, S. H, Henrickson, S. E, von Andrian, U. H, Shlomchik, M. J, & Haberman, A. M. (2007) Definition of germinal-center B cell migration in vivo reveals predominant intrazonal circulation patterns. *Immunity* **26**, 655–667.

[25] Schwickert, T. A, Lindquist, R. L, Shakhar, G, Livshits, G, Skokos, D, Kosco-Vilbois, M. H, Dustin, M. L, & Nussenzweig, M. C. (2007) In vivo imaging of germinal centres reveals a dynamic open structure. *Nature* **446**, 83–87.

[26] Allen, C. D, Okada, T, & Cyster, J. G. (2007) Germinal-center organization and cellular dynamics. *Immunity* **27**, 190–202.

[27] Bernard, S & Herzel, H. (2006) Why do cells cycle with a 24 hour period? *Genome informatics. International Conference on Genome Informatics* **17**, 72–79.

[28] Sakaue-Sawano, A, Kurokawa, H, Morimura, T, Hanyu, A, Hama, H, Osawa, H, Kashiwagi, S, Fukami, K, Miyata, T, Miyoshi, H, Imamura, T, Ogawa, M, Masai, H, & Miyawaki, A. (2008) Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498.

[29] Hama, H, Kurokawa, H, Kawano, H, Ando, R, Shimogori, T, Noda, H, Fukami, K, Sakaue-Sawano, A, & Miyawaki, A. (2011) Scale: a chemical approach for fluorescence imaging and reconstruction of transparent mouse brain. *Nat Neurosci* **14**, 1481–1488.

[30] Hlatky, L, Olesiak, M, & Hahnfeldt, P. (1996) Measurement of Potential Doubling Time for Human Tumor Xenografts Using the Cytokinesis-Block Method. *Cancer Research* **56**, 1660–1663.

[31] Bourhis, J, Dendale, R, Hill, C, Bosq, J, Janot, F, Attal, P, Fortin, A, Marandas, P, Schwaab, G, Wibault, P, Malaise, E. P, Bobin, S, Luboinski, B, Eschwege, F, & Wilson, G. (1996) Potential doubling time and clinical outcome in head and neck squamous cell carcinoma treated with 70 GY in 7 weeks. *International journal of radiation oncology, biology, physics* **35**, 471–476.

[32] Dubray, B, Maciorowsky, Z, Cosset, J. M, & Terry, N. H. (1995) [Clinical value of the potential doubling time (Tpot) measured by flow cytometry]. *Bulletin du cancer* **82**, 331–338.

[33] Pollack, A, Terry, N. H, Wu, C. S, Wise, B. M, White, R. A, & Meistrich, M. L. (1995) Specific staining of iododeoxyuridine and bromodeoxyuridine in tumors double labelled in vivo: a cell kinetic analysis. *Cytometry* **20**, 53–61.

[34] Janeway, C, Murphy, K. P, Travers, P, Walport, M, & Janeway, C. (2008) *Janeway's immuno biology.* (Garland Science).

[35] Mendelsohn, M. L. (1962) Autoradiographic Analysis of Cell Proliferation in Spontaneous Breast Cancer of C3H Mouse. III. The Growth Fraction. *Journal of the National Cancer Institute* **28**, 1015–1029.

[36] Zhang, J, MacLennan, I. C, Liu, Y. J, & Lane, P. J. (1988) Is rapid proliferation in B centroblasts linked to somatic mutation in memory B cell clones? *Immunology letters* **18**, 297–299.

[37] Allen, C. D. C, Okada, T, Tang, H. L, & Cyster, J. G. (2007) Imaging of Germinal Center Selection Events During Affinity Maturation. *Science* **315**, 528–531.

[38] Zaitoun, A. M. (1980) Cell population kinetics of the germinal centres of lymph nodes of BALB/c mice. *Journal of Anatomy* **130**, 131–137.

[39] León, K, Faro, J, & Carneiro, J. (2004) A general mathematical framework to model generation structure in a population of asynchronously dividing cells. *Journal of Theoretical Biology* **229**, 455–476.

[40] Eisen, M. B, Spellman, P. T, Brown, P. O, & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

[41] Puck, T. T, Marcus, P. I, & Cieciura, S. J. (1956) Clonal growth of mammalian cells in vitro. *The Journal of Experimental Medicine* **103**, 273–284.

[42] Hawkins, E. D, Markham, J. F, McGuinness, L. P, & Hodgkin, P. D. (2009) A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 13457–13462.

[43] Mohri, H, Bonhoeffer, S, Monard, S, Perelson, A. S, & Ho, D. D. (1998) Rapid turnover of T lymphocytes in SIV-infected rhesus macaques. *Science* **279**, 1223–1227.

[44] Venter, J. C, Adams, M. D, Myers, E. W, Li, P. W, & Mural, R. J. (2001) The Sequence of the Human Genome. *Science* **291**, 1304–1351.

[45] Kunkel, T. A. (2004) DNA Replication Fidelity. *Journal of Biological Chemistry* **279**, 16895–16898.

[46] Ng, R. K & Gurdon, J. B. (2008) Epigenetic inheritance of cell differentiation status. *Cell cycle (Georgetown, Tex.)* **7**, 1173–1177.

[47] Tyson, J. J & Novak, B. (2008) Temporal organization of the cell cycle. *Current Biology* **18**, R759–R768.

[48] Abraham, R. T. (2001) Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes & Development* **15**, 2177–2196.

[49] Zimmermann, K. C & Green, D. R. (2001) How cells die: apoptosis pathways. *The Journal of allergy and clinical immunology* **108**.

[50] Nasmyth, K. (1996) Viewpoint: Putting the Cell Cycle in Order. *Science* **274**, 1643–1645.

[51] Hübscher, U, Maga, G, & Spadari, S. (2002) Eukaryotic dna polymerases. *Annual Review of Biochemistry* **71**, 133–163.

[52] Sullivan, M & Morgan, D. O. (2007) Finishing mitosis, one step at a time. *Nature Reviews Molecular Cell Biology* **8**, 894–903.

[53] Vaithiyalingam, S, Warren, E. M, Eichman, B. F, & Chazin, W. J. (2010) Insights into eukaryotic DNA priming from the structure and functional interactions of the 4Fe-4S cluster domain of human DNA primase. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 13684–13689.

[54] Russell, P & Nurse, P. (1986) Schizosaccharomyces pombe and Saccharomyces cerevisiae: a look at yeasts divided. *Cell* **45**, 781–782.

[55] Evans, T, Rosenthal, E. T, Youngblom, J, Distel, D, & Hunt, T. (1983) Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. *Cell* **33**, 389–396.

[56] Nurse, P. (1990) Universal control mechanism regulating onset of M-phase. *Nature* **344**, 503–508.

[57] Bloom, J & Cross, F. R. (2007) Multiple levels of cyclin specificity in cell-cycle control. *Nat Rev Mol Cell Biol* **8**, 149–160.

[58] Morgan, D. O. (1995) Principles of CDK regulation. *Nature* **374**, 131–134.

[59] Flemming, W. (1879) Beitrage zur Kenntniss der Zelle und ihrer Lebenserscheinungen. **16**, 302–436.

[60] Hollowood, K & Macartney, J. (1992) Cell kinetics of the germinal center reaction–a stathmokinetic study. *European Journal of Immunology* **22**, 261–266.

[61] Hanna, M. G. (1964) An autoradiographic study of the germinal center in spleen white pulp during early intervals of the immune response. *Laboratory Investigation* **13**, 95–104.

[62] Dolbeare, F & Selden, J. R. (1994) Immunochemical quantitation of bromodeoxyuridine: application to cell-cycle kinetics. *Methods in cell biology* **41**, 297–316.

[63] Ribeiro, R. M. (2007) Dynamics of CD4(+) T cells in HIV-1 infection. *Immunology and Cell Biology* **85**, 287–294.

[64] Ganusov, V. V, Borghans, J. A, & De Boer, R. J. (2010) Explicit kinetic heterogeneity: mathematical models for interpretation of deuterium labeling of heterogeneous cell populations. *PLoS Computational Biology* **6**, e1000666+.

[65] De Boer, R. J, Ganusov, V. V, Milutinović, D, Hodgkin, P. D, & Perelson, A. S. (2006) Estimating lymphocyte division and death rates from CFSE data. *Bulletin of Mathematical Biology* **68**, 1011–1031.

[66] Lyons, A. B. (2000) Analysing cell division in vivo and in vitro using flow cytometric measurement of CFSE dye dilution. *Journal of immunological methods* **243**, 147–154.

[67] Oostendorp, R. A. J, Audet, J, & Eaves, C. J. (2000) High-resolution tracking of cell division suggests similar cell cycle kinetics of hematopoietic stem cells stimulated in vitro and in vivo. *Blood* **95**, 855–862.

[68] Spellman, P. T, Sherlock, G, Zhang, M. Q, Iyer, V. R, Anders, K, Eisen, M. B, Brown, P. O, Botstein, D, & Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell* **9**, 3273–3297.

[69] Shedden, K & Cooper, S. (2002) Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 4379–4384.

[70] Cooper, S & Shedden, K. (2003) Microarray analysis of gene expression during the cell cycle. *Cell & chromosome* **2**, 1+.

[71] Hoeijmakers, W. A, Bártfai, R, & Stunnenberg, H. G. (2013) Transcriptome analysis using RNA-Seq. *Methods in molecular biology (Clifton, N.J.)* **923**, 221–239.

[72] Hahn, A. T, Jones, J. T, & Meyer, T. (2009) Quantitative analysis of cell cycle phase durations and PC12 differentiation using fluorescent biosensors. *Cell Cycle* **8**, 1044–1052.

[73] Winsor, C. P. (1932) The Gompertz Curve as a Growth Curve. *Proceedings of the National Academy of Sciences of the United States of America* **18**, 1–8.

[74] Smith, J. A, Laurence, D. J, & Rudland, P. S. (1981) Limitations of cell kinetics in distinguishing cell cycle models. *Nature* **293**, 648–650.

[75] Nelson, S & Green, P. J. (1981) The random transition model of the cell cycle. A critical review. *Cancer Chemotherapy and Pharmacology* **6**, 11–18.

[76] Tyrcha, J. (2001) Age-dependent cell cycle models. *Journal of Theoretical Biology* **213**, 89–101.

[77] Cain, S. J & Chau, P. C. (1997) A transition probability cell cycle model simulation of bivariate DNA/bromodeoxyuridine distributions. *Cytometry* **27**, 239–249.

[78] Lee, H. Y. Y & Perelson, A. S. (2008) Modeling T cell proliferation and death in vitro based on labeling data: generalizations of the Smith-Martin cell cycle model. *Bulletin of mathematical biology* **70**, 21–44.

[79] Ganusov, V. V, Pilyugin, S. S, de Boer, R. J, Murali-Krishna, K, Ahmed, R, & Antia, R. (2005) Quantifying cell turnover using CFSE data. *Journal of Immunological Methods* **298**, 183–200.

[80] Swierniak, A, Polanski, A, & Kimmel, M. (1996) Optimal control problems arising in cell-cycle-specific cancer chemotherapy. *Cell proliferation* **29**, 117–139.

[81] Basse, B & Ubezio, P. (2007) A generalised age- and phase-structured model of human tumour cell populations both unperturbed and exposed to a range of cancer therapies. *Bulletin of mathematical biology* **69**, 1673–1690.

[82] Bernard, S, Čajavec Bernard, B, Lévi, F, & Herzel, H. (2010) Tumor Growth Rate Determines the Timing of Optimal Chronomodulated Treatment Schedules. *PLoS Comput Biol* **6**, e1000712+.

[83] Altinok, A, Lévi, F, & Goldbeter, A. (2009) Identifying mechanisms of chronotolerance and chronoefficacy for the anticancer drugs 5-fluorouracil and oxaliplatin by computational modeling. *European Journal of Pharmaceutical Sciences* **36**, 20–38.

[84] Swierniak, A, Kimmel, M, & Smieja, J. (2009) Mathematical modeling as a tool for planning anticancer therapy. *European journal of pharmacology* **625**, 108–121.

[85] Meyer-Hermann, M, Mohr, E, Pelletier, N, Zhang, Y, Victora, G. D, & Toellner, K.-M. M. (2012) A theory of germinal center B cell selection, division, and exit. *Cell reports* **2**, 162–174.

[86] Zilman, A, Ganusov, V. V, & Perelson, A. S. (2010) Stochastic models of lymphocyte proliferation and death. *PloS One* **5**, e12775+.

[87] Altinok, A, Gonze, D, Lévi, F, & Goldbeter, A. (2011) An automaton model for the cell cycle. *Interface focus* **1**, 36–47.

[88] Subramanian, V. G, Duffy, K. R, Turner, M. L, & Hodgkin, P. D. (2008) Determining the expected variability of immune responses using the cyton model. *Journal of mathematical biology* **56**, 861–892.

[89] De Boer, R. J, Mohri, H, Ho, D. D, & Perelson, A. S. (2003) Estimating average cellular turnover from 5-bromo-2'-deoxyuridine (BrdU) measurements. *Proceedings Biological Sciences* **270**, 849–858.

[90] Johnson, N. L. (1960) An Approximation to the Multinomial Distribution: Some Properties and Applications. *Biometrika* **47**, 93+.

[91] Roberts, G. O & Rosenthal, J. S. (2009) Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* **18**, 349–367.

[92] Bel, G, Munsky, B, & Nemenman, I. (2009) The simplicity of completion time distributions for common complex biochemical processes. *Physical Biology* **7**, 016003+.

[93] Duffy, K. R, Wellard, C. J, Markham, J. F, Zhou, J. H. S, Holmberg, R, Hawkins, E. D, Hasbold, J, Dowling, M. R, & Hodgkin, P. D. (2012) Activation-Induced B Cell Fates Are Selected by Intracellular Stochastic Competition. *Science* **335**, 338–341.

[94] Allman, E. S, Matias, C, & Rhodes, J. A. (2009) Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics* **37**, 3099–3132.

[95] Rothenberg, T. J. (1971) Identification in Parametric Models. *Econometrica* **39**, 577+.

[96] Stigler, S. M. (1974) Gergonne's 1815 paper on the design and analysis of polynomial regression experiments. *Historia Mathematica* **1**, 431–447.

[97] Atkinson, A. C & Bogacka, B. (2012) Optimum designs for the equality of parameters in enzyme inhibition kinetic models. *Journal of Statistical Planning and Inference*.

[98] Jennrich, R. I. (1969) Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics* **40**, 633–643.

[99] Wald, A. (1943) On the Efficient Design of Statistical Investigations. *The Annals of Mathematical Statistics* **14**, 134–140.

[100] Box, G. E. P & Lucas, H. L. (1959) Design of Experiments in Non-Linear Situations. *Biometrika* **46**, 77+.

[101] Draper, N. R & Hunter, W. G. (1966) Design of Experiments for Parameter Estimation in Multiresponse Situations. *Biometrika* **53**, 525+.

[102] Atkinson, A. C & Donev, A. N. (1992) *Optimum Experimental Designs (Oxford Statistical Science Series)*. (Oxford University Press, USA).

[103] Dette, H & O'Brien, T. E. (1999) Optimality criteria for regression models based on predicted variance. *Biometrika* **86**, 93–106.

[104] Chaloner, K & Verdinelli, I. (1995) Bayesian Experimental Design: A Review. *Statistical Science* **10**, 273–304.

[105] Gautier, R & Pronzato, L. (1998) *Sequential Design and Active Control*. (Institute of Mathematical Statistics, Hayward, CA), pp. 138–151.

[106] Wynn, H. P. (1970) The Sequential Generation of D-Optimum Experimental Designs. *The Annals of Mathematical Statistics* **41**, 1655–1664.

[107] Atwood, C. L. (1973) Sequences Converging to D-Optimal Designs of Experiments. *The Annals of Statistics* **1**, 342–352.

[108] Fedorov, V. V. (1971) The Design of Experiments in the Multiresponse Case. *Theory of Probability and its Applications* **16**, 323–332.

[109] Gratzner, H. G. (1982) Monoclonal antibody to 5-bromo- and 5-iododeoxyuridine: A new reagent for detection of DNA replication. *Science (New York, N.Y.)* **218**, 474–475.

[110] Thornton, J. G, Wells, M, & Hume, W. J. (1988) Flash labelling of S-phase cells in short-term organ culture of normal and pathological human endometrium using bromodeoxyuridine and tritiated thymidine. *The Journal of pathology* **154**, 321–328.

[111] Hyatt, G. A & Beebe, D. C. (1992) Use of a double-label method to detect rapid changes in the rate of cell proliferation. *The Journal of Histochemistry and Cytochemistry* **40**, 619–627.

[112] Bakker, P. J, Stap, J, Tukker, C. J, van Oven, C. H, Veenhof, C. H, & Aten, J. (1991) An indirect immunofluorescence double staining procedure for the simultaneous flow cytometric measurement of iodo- and chlorodeoxyuridine incorporated into DNA. *Cytometry* **12**, 366–372.

[113] Aten, J. A, Bakker, P. J. M, Stap, J, Boschman, G. A, & Veenhof, C. H. N. (1992) DNA double labelling with IdUrd and CldUrd for spatial and temporal analysis of cell proliferation and DNA replication. **24**, 251–259.

[114] Buck, S. B, Bradford, J, Gee, K. R, Agnew, B. J, Clarke, S. T, & Salic, A. (2008) Detection of S-phase cell cycle progression using 5-ethynyl-2'-deoxyuridine incorporation with click chemistry, an alternative to using 5-bromo-2'-deoxyuridine antibodies. *BioTechniques* **44**, 927–929.

[115] Salic, A & Mitchison, T. J. (2008) A chemical method for fast and sensitive detection of DNA synthesis in vivo. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 2415–2420.

[116] Bradford, J. A & Clarke, S. T. (2001) Dual-Pulse Labeling Using 5-Ethynyl-2-Deoxyuridine (EdU) and 5-Bromo-2-Deoxyuridine (BrdU) in Flow Cytometry.

[117] Lentz, S. I, Edwards, J. L, Backus, C, McLean, L. L, Haines, K. M, & Feldman, E. L. (2010) Mitochondrial DNA (mtDNA) Biogenesis: Visualization and Duel Incorporation of BrdU and EdU Into Newly Synthesized mtDNA In Vitro. *The journal of histochemistry and cytochemistry* **58**, 207–218.

[118] Igel, C, Hansen, N, & Roth, S. (2007) Covariance matrix adaptation for multi-objective optimization. *Evolutionary computation* **15**, 1–28.

[119] Griewank, A, Juedes, D, Mitev, H, Utke, J, Vogel, O, & Walther, A. (1999) ADOL-C: A Package for the Automatic Differentiation of Algorithms Written in C/C++.

[120] Bates, D. M & Watts, D. G. (2007) *Nonlinear Regression Analysis and Its Applications (Wiley Series in Probability and Statistics)*. (Wiley-Interscience).

[121] Dempster, A. P, Laird, N. M, & Rubin, D. B. (1977) *Maximum likelihood from incomplete data via the EM algorithm*. Vol. 39, pp. 1–38.

[122] Igel, C, Heidrich-Meisner, V, & Glasmachers, T. (2008) Shark. *Journal of Machine Learning Research* **9**, 993–996.

[123] Miguel. (2000) Mode-Finding for Mixtures of Gaussian Distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **22**, 1318–1323.

[124] Channel, S. R & Hancock, B. L. (1993) Application of kinetic models to estimate transit time through cell cycle compartments. *Toxicology letters* **68**, 213–221.

[125] Levkoff, L. H, Marshall, G. P, Ross, H. H, Caldeira, M, Reynolds, B. A, Cakiroglu, M, Mariani, C. L, Streit, W. J, & Laywell, E. D. (2008) Bromodeoxyuridine inhibits cancer cell proliferation in vitro and in vivo. *Neoplasia (New York, N.Y.)* **10**, 804–816.

[126] Tuttle, A. H, Rankin, M. M, Teta, M, Sartori, D. J, Stein, G. M, Kim, G. J, Virgilio, C, Granger, A, Zhou, D, Long, S. H, Schiffman, A. B, & Kushner, J. A. (2010) Immunofluorescent Detection of Two Thymidine Analogues (CldU and IdU) in Primary Tissue. *Journal of Visualized Experiments* pp. e2166+.

[127] Wu, Y, Guo, F, Liu, J, Xiao, X, Huang, L, & He, D. (2011) Triple labeling with three thymidine analogs reveals a well-orchestrated regulation of hepatocyte proliferation during liver regeneration. *Hepatology Research* **41**, 1230–1239.

[128] Batista, F. D & Harwood, N. E. (2009) The who, how and where of antigen presentation to B cells. *Nature reviews. Immunology* **9**, 15–27.

[129] Okada, T, Miller, M. J, Parker, I, Krummel, M. F, Neighbors, M, Hartley, S. B, O'Garra, A, Cahalan, M. D, & Cyster, J. G. (2005) Antigen-Engaged B Cells Undergo Chemotaxis toward the T Zone and Form Motile Conjugates with Helper T Cells. *PLoS Biol* **3**, e150+.

[130] Garside, P, Ingulli, E, Merica, R. R, Johnson, J. G, Noelle, R. J, & Jenkins, M. K. (1998) Visualization of Specific B and T Lymphocyte Interactions in the Lymph Node. *Science* **281**, 96–99.

[131] Steinman, R. M, Inaba, K, Turley, S, Pierre, P, & Mellman, I. (1999) Antigen capture, processing, and presentation by dendritic cells: recent cell biological studies. *Human immunology* **60**, 562–567.

[132] McHeyzer Williams, L. J & McHeyzer Williams, M. G. (2005) Antigen-specific memory b cell development. *Annual Review of Immunology* **23**, 487–513.

[133] Jacob, J & Kelsoe, G. (1992) In situ studies of the primary immune response to (4-hydroxy-3-nitrophenyl)acetyl. II. A common clonal origin for periarteriolar lymphoid sheath-associated foci and germinal centers. *The Journal of experimental medicine* **176**, 679–687.

[134] Liu, Y.-J, Zhang, J, Lane, P. J. L, Chan, E. Y. T, & Maclennan, I. C. M. (1991) Sites of specific B cell activation in primary and secondary responses to T cell-dependent and T cell-independent antigens. *Eur. J. Immunol.* **21**, 2951–2962.

[135] Camacho, S. A, Kosco-Vilbois, M. H, & Berek, C. (1998) The dynamic structure of the germinal center. *Immunology today* **19**, 511–514.

[136] Eisen, H. N & Siskind, G. W. (1964) Variations in Affinities of Antibodies during the Immune Response*. *Biochemistry* **3**, 996–1008.

[137] Tarlinton, D. M & Smith, K. G. (2000) Dissecting affinity maturation: a model explaining selection of antibody-forming cells and memory B cells in the germinal centre. *Immunology today* **21**, 436–441.

[138] Tarlinton, D. M. (2008) Evolution in miniature: selection, survival and distribution of antigen reactive cells in the germinal centre. *Immunology and cell biology* **86**, 133–138.

[139] Flemming, W. (1884) Studien über Regeneration der Gewebe. **24**, 50–91.

[140] Nieuwenhuis, P & Opstelten, D. (1984) Functional anatomy of germinal centers. *The American journal of anatomy* **170**, 421–435.

[141] Victora, G. D, Dominguez-Sola, D, Holmes, A. B, Deroubaix, S, Dalla-Favera, R, & Nussenzweig, M. C. (2012) Identification of human germinal center light and dark zone cells and their relationship to human B-cell lymphomas. *Blood* **120**, 2240–2248.

[142] Berek, C, Berger, A, & Apel, M. (1991) Maturation of the immune response in germinal centers. *Cell* **67**, 1121–1129.

[143] Berek, C & Milstein, C. (1987) Mutation Drift and Repertoire Shift in the Maturation of the Immune Response. *Immunological Reviews* **96**, 23–41.

[144] Ziegner, M, Steinhauser, G, & Berek, C. (1994) Development of antibody diversity in single germinal centers: selective expansion of high-affinity variants. *European journal of immunology* **24**, 2393–2400.

[145] Kleinstein, S. H & Singh, J. P. P. (2003) Why are there so few key mutant clones? The influence of stochastic selection and blocking on affinity maturation in the germinal center. *International immunology* **15**, 871–884.

[146] Good-Jacobson, K. L & Shlomchik, M. J. (2010) Plasticity and heterogeneity in the generation of memory B cells and long-lived plasma cells: the influence of germinal center interactions and dynamics. *Journal of immunology (Baltimore, Md. : 1950)* **185**, 3117–3125.

[147] Klein, U & Dalla-Favera, R. (2008) Germinal centres: role in B-cell physiology and malignancy. *Nature Reviews Immunology* **8**, 22–33.

[148] Radbruch, A, Muehlinghaus, G, Luger, E. O, Inamine, A, Smith, K. G, Dörner, T, & Hiepe, F. (2006) Competence and competition: the challenge of becoming a long-lived plasma cell. *Nature reviews. Immunology* **6**, 741–750.

[149] Zotos, D & Tarlinton, D. M. (2012) Determining germinal centre B cell fate. *Trends in immunology* **33**, 281–288.

[150] Giang, T, Green, J. A, Gray, E. E, Xu, Y, & Cyster, J. G. (2009) Immune complex relay by subcapsular sinus macrophages and noncognate B cells drives antibody affinity maturation. *Nature Immunology* **10**, 786–793.

[151] Muramatsu, M, Kinoshita, K, Fagarasan, S, Yamada, S, Shinkai, Y, & Honjo, T. (2000) Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563.

[152] MacLennan, I. C. (1994) Germinal centers. *Annual Review of Immunology* **12**, 117–139.

[153] Victora, G. D, Schwickert, T. A, Fooksman, D. R, Kamphorst, A. O, Meyer-Hermann, M, Dustin, M. L, & Nussenzweig, M. C. (2010) Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter. *Cell* **143**, 592–605.

[154] Kepler, T. B & Perelson, A. S. (1993) Somatic hypermutation in B cells: an optimal control treatment. *Journal of theoretical biology* **164**, 37–64.

[155] Oprea, M, van Nimwegen, E, & Perelson, A. S. (2000) Dynamics of one-pass germinal center models: implications for affinity maturation. *Bulletin of mathematical biology* **62**, 121–153.

[156] Kleinstein, S. H & Singh, J. P. (2001) Toward quantitative simulation of germinal center dynamics: biological and modeling insights from experimental validation. *Journal of theoretical biology* **211**, 253–275.

[157] Meyer-Hermann, M. E, Maini, P. K, & Iber, D. (2006) An analysis of B cell selection mechanisms in germinal centers. *Mathematical medicine and biology : a journal of the IMA* **23**, 255–277.

[158] Oprea, M & Perelson, A. S. (1997) Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts. *Journal of immunology (Baltimore, Md. : 1950)* **158**, 5155–5162.

[159] Radmacher, M. D, Kelsoe, G, & Kepler, T. B. (1998) Predicted and inferred waiting times for key mutations in the germinal centre reaction: evidence for stochasticity in selection. *Immunology and cell biology* **76**, 373–381.

[160] Figge, M. T. T. (2005) Stochastic discrete event simulation of germinal center reactions. *Physical review. E, Statistical, nonlinear, and soft matter physics* **71**.

[161] Iber, D & Maini, P. K. (2002) A mathematical model for germinal centre kinetics and affinity maturation. *Journal of theoretical biology* **219**, 153–175.

[162] Keşmir, C & De Boer, R. J. (2003) A spatial model of germinal center reactions: cellular adhesion based sorting of B cells results in efficient affinity maturation. *Journal of theoretical biology* **222**, 9–22.

[163] Wittenbrink, N, Weber, T. S, Klein, A, Weiser, A. A, Zuschratter, W, Sibila, M, Schuchhardt, J, & Or-Guil, M. (2010) Broad Volume Distributions Indicate Nonsynchronized Growth and Suggest Sudden Collapses of Germinal Center B Cell Populations. *The Journal of Immunology* **184**, 1339–1347.

[164] Keşmir, C & De Boer, R. J. (1999) A mathematical model on germinal center kinetics and termination. *Journal of immunology (Baltimore, Md. : 1950)* **163**, 2463–2469.

[165] Moreira, J. S & Faro, J. (2006) Modelling two possible mechanisms for the regulation of the germinal center dynamics. *Journal of immunology (Baltimore, Md. : 1950)* **177**, 3705–3710.

[166] Hawkins, J. B, Jones, M. T, Plassmann, P. E, & Thorley-Lawson, D. A. (2011) Chemotaxis in Densely Populated Tissue Determines Germinal Center Anatomy and Cell Motility: A New Paradigm for the Development of Complex Tissues. *PLoS ONE* **6**, e27650+.

[167] Wittenbrink, N, Klein, A, Weiser, A. A, Schuchhardt, J, & Or-Guil, M. (2011) Is there a typical germinal center? A large-scale immunohistological study on the cellular composition of germinal centers during the hapten-carrier-driven primary immune response in mice. *Journal of Immunology* **187**, 6185–6196.

[168] Beltman, J. B, Allen, C. D. C, Cyster, J. G, & de Boer, R. J. (2011) B cells within germinal centers migrate preferentially from dark to light zone. *Proceedings of the National Academy of Sciences* **108**, 8755–8760.

[169] Kerfoot, S. M, Yaari, G, Patel, J. R, Johnson, K. L, Gonzalez, D. G, Kleinstein, S. H, & Haberman, A. M. (2011) Germinal Center B Cell and T Follicular Helper Cell Development Initiates in the Interfollicular Zone. *Immunity* **34**, 947–960.

[170] Khalil, A. M, Cambier, J. C, & Shlomchik, M. J. (2012) B cell receptor signal transduction in the GC is short-circuited by high phosphatase activity. *Science (New York, N.Y.)* **336**, 1178–1181.

[171] Méndez-Ferrer, S, Michurina, T. V, Ferraro, F, Mazloom, A. R, Macarthur, B. D, Lira, S. A, Scadden, D. T, Ma'ayan, A, Enikolopov, G. N, & Frenette, P. S. (2010) Mesenchymal and haematopoietic stem cells form a unique bone marrow niche. *Nature* **466**, 829–834.

[172] Allen, C. D, Ansel, K. M, Low, C, Lesley, R, Tamamura, H, Fujii, N, & Cyster, J. G. (2004) Germinal center dark and light zone organization is mediated by CXCR4 and CXCR5. *Nature immunology* **5**, 943–952.

[173] Barberis, M, Beck, C, Amoussouvi, A, Schreiber, G, Diener, C, Herrmann, A, & Klipp, E. (2011) A low number of SIC1 mRNA molecules ensures a low noise level in cell cycle progression of budding yeast. *Molecular bioSystems* **7**, 2804–2812.

[174] Kar, S, Baumann, W. T, Paul, M. R, & Tyson, J. J. (2009) Exploring the roles of noise in the eukaryotic cell cycle. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6471–6476.

[175] Thattai, M & van Oudenaarden, A. (2001) Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 8614–8619.

[176] den Iseger, P. (2006) Numerical transform inversion using gaussian quadrature. *Probability in the Engineering and Informational Sciences* **20**, 1–44.

[177] Zetterberg, A & Larsson, O. (1985) Kinetic analysis of regulatory events in G1 leading to proliferation or quiescence of Swiss 3T3 cells. *Proceedings of the National Academy of Sciences of the United States of America* **82**, 5365–5369.

[178] Tinnemans, M. M, Schutte, B, Lenders, M. H, Ten Velde, G. P, Ramaekers, F. C, & Blijham, G. H. (1993) Cytokinetic analysis of lung cancer by in vivo bromodeoxyuridine labelling. *British journal of cancer* **67**, 1217–1222.

[179] Groves, M. D, Maor, M. H, Meyers, C, Kyritsis, A. P, Jaeckle, K. A, Yung, Sawaya, R. E, Hess, K, Bruner, J. M, Peterson, P, & Levin, V. A. (1999) A phase II trial of high-dose bromodeoxyuridine with accelerated fractionation radiotherapy followed by procarbazine, lomustine, and vincristine for glioblastoma multiforme. *International Journal of Radiation Oncology*Biology*Physics* **45**, 127–135.

[180] Robertson, J. M, McGinn, C. J, Walker, S, Marx, Kessler, M. L, Ensminger, W. D, & Lawrence, T. S. (1997) A Phase I trial of hepatic arterial bromodeoxyuridine and conformal radiation therapy for patients with primary hepatobiliary cancers or colorectal liver metastases. *International Journal of Radiation Oncology*Biology*Physics* **39**, 1087–1092.

[181] Srinivasula, S, Lempicki, R. A, Adelsberger, J. W, Huang, C.-Y, Roark, J, Lee, P. I, Rupert, A, Stevens, R, Sereti, I, Lane, H. C, Di Mascio, M, & Kovacs, J. A. (2011) Differential effects of HIV viral load and CD4 count on proliferation of naive and memory CD4 and CD8 T lymphocytes. *Blood* **118**, 262–270.

[182] Collins, M. L, Eng, S, Hoh, R, & Hellerstein, M. K. (2003) Measurement of mitochondrial DNA synthesis in vivo using a stable isotope-mass spectrometric technique. *Journal of applied physiology (Bethesda, Md. : 1985)* **94**, 2203–2211.

[183] Borghans, J. A & de Boer, R. J. (2007) Quantification of T-cell dynamics: from telomeres to DNA labeling. *Immunological reviews* **216**, 35–47.

[184] Larsson, S, Johansson, M, Oredsson, S, & Holst, U. (2005) A Markov model approach shows a large variation in the length of S phase in MCF-7 breast cancer cells. *Cytometry* **65A**, 15–25.

[185] Maler, A & Lutscher, F. (2010) Cell-cycle times and the tumour control probability. *Mathematical medicine and biology : a journal of the IMA* **27**, 313–342.

[186] Yousuf, N, Yanik, G. A, George, B. A, Masterson, M, Mazewski, C. M, White, L. M, Miller, M. A, Lampkin, B. C, & Raza, A. (1991) Comparison of two double labeling techniques to measure cell cycle kinetics in myeloid leukemias. *Anticancer research* **11**, 1195–1199.

[187] Begg, A. C. (1993) Critical Appraisal of In Situ Cell Kinetic Measurements as Response Predictors in Human Tumors. *Seminars in radiation oncology* **3**, 144–151.

[188] Shirakawa, S, Luce, J. K, Tannock, I, & Frei, E. (1970) Cell proliferation in human melanoma. *The Journal of clinical investigation* **49**, 1188–1199.

[189] Risch, R. H. (1969) The Problem of Integration in Finite Terms. *Transactions of the American Mathematical Society* **139**, 167+.

[190] Figueiredo, M. A. T & Jain, A. K. (2002) Unsupervised learning of finite mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **24**, 381–396.

[191] Chib, S & Greenberg, E. (1995) Understanding the Metropolis-Hastings Algorithm. *The American Statistician* **49**, 327–335.

[192] Metropolis, N, Rosenbluth, A. W, Rosenbluth, M. N, Teller, A. H, & Teller, E. (1953) Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics* **21**, 1087–1092.

[193] Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

[194] Haario, H, Saksman, E, & Tamminen, J. (2001) An Adaptive Metropolis Algorithm. *Bernoulli* **7**, 223+.

[195] Cai, B, Meyer, R, & Perron, F. (2008) Metropolis-Hastings algorithms with adaptive proposals. **18**, 421–433.

[196] Lapeyre, B & Lelong, J. (2010) A framework for adaptive Monte-Carlo procedures. *Monte Carlo Methods and Applications* **17**.

[197] Hansen, N. (2005) The CMA Evolution Strategy: A Tutorial.

[198] Arthur, D & Vassilvitskii, S. (2007) *k-means++: the advantages of careful seeding*, SODA '07. (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA), pp. 1027–1035.

[199] Lo, K, Brinkman, R. R. R, & Gottardo, R. (2008) Automated gating of flow cytometry data via robust model-based clustering. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* **73**, 321–332.

[200] Lee, G & Scott, C. (2012) EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Computational Statistics & Data Analysis* **56**, 2816–2829.

[201] Ge, Y & Sealfon, S. C. (2012) flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics* **28**, 2052–2058.

List of Figures

List of Tables